INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2025IPPAT016

Thèse de doctorat

From trustworthy AI to technical standards – The distinctive European approach to artificial intelligence regulation

Thèse de doctorat de l’Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l’Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Droit et Régulation

Thèse présentée et soutenue à Paris, le 19 Mai 2025, par

MÉLANIE GORNET

Composition du Jury :

Jean-Gabriel Ganascia Professeur émérite, Sorbonne Université	Président / Examinateur
Céline Castets-Renard Professeure, Université d’Ottawa	Rapporteure
Fabien Tarissan Chargé de recherche, CNRS, ENS Paris-Saclay	Rapporteur
Irene Kamara Assistant Professor, Tilburg Law School	Examinatrice
Célia Zolynski Professeure, Université Paris 1 Panthéon-Sorbonne	Examinatrice
Winston Maxwell Professeur, Télécom Paris (i3)	Directeur de thèse
Florence d’Alché-Buc Professeure, Télécom Paris (LTCI)	Co-directrice de thèse
Tiphaine Viard Maîtresse de Conférence, Télécom Paris (i3)	Co-encadrante de thèse

The views and opinions expressed in this thesis are those of the author alone. They do not reflect the views of Télécom Paris, nor the doctoral school, nor those of the people who were interviewed or who collaborated with the author. Finally, the contents of this thesis do not necessarily reflect the position or opinion of the European Commission.

This thesis was written in November 2024. It has been updated in March 2025.
Main updates are presented in boxes throughout the manuscript.

Abstract

Europe has been at the forefront of Artificial Intelligence (AI) ethics, developing non-binding charters and principles on "trustworthy" AI. The term "trustworthiness" is used by Europe to designate AI systems that are "ethical", "legal" and "technically robust". Europe has supplemented these non-binding principles with a binding regulation on AI, known as the AI Act. The AI Act is one of the world's first comprehensive frameworks for regulating AI systems across different industries and use cases, focusing on safety and protection of fundamental rights. The AI Act relies, for operational questions, mostly on technical standards that are in the course of development. The European approach thus combines three layers of regulatory instruments: AI ethics charters, the AI Act and technical standards. These three regulatory layers are meant to be complementary, but their interplay is challenging. Indeed, translating ethical principles into legal obligations and finally into technical criteria is not straightforward. In this manuscript, we therefore ask:

What makes Europe's approach to regulating AI so distinctive from other European regulations?

We first look at each of these regulatory mechanisms, how they work, independently and together. First, through a quantitative analysis of 436 AI ethics charters, we show how some AI ethics principles remain specific to certain categories of stakeholders, while others have fully permeated the space of AI ethics. We also highlight the underlining power dynamics around AI ethics, where charters are used as a means of communication and influence.

The AI Act takes roots in some of these ethics charters, as it was notably influenced by the work of the High Level Expert Group on AI (HLEG) set out by the European Commission. But the AI Act also builds upon various other European frameworks, including the General Data Protection Regulation (GDPR) and product safety regulations. Technical standards, however, have historically remained separate from the legislation. Instead, they served as an economic tool to shape market dynamics while ensuring products interoperability and quality. In recent years, they have taken on a new role in Europe under the New Legislative Framework (NLF), where technical "harmonised" standards are used to ensure compliance with legal requirements. This is the approach that Europe chose for the AI Act.

This study of regulatory mechanisms, shows that many normative choices are in fact made at the end of the line, in technical standards. The standardisation approach is frequent in product safety, but under the AI Act, standards are also expected to address fundamental rights concerns. Even when defining seemingly technical terms and properties, standards are actually value-laden. The feasibility of developing technical requirements on criteria such as fairness, inherited from ethics charters, is therefore difficult. To avoid making hard normative choices, stan-

dardisation organisations are playing it safe, developing standards which remain at a high-level. As a result, standards loop back to ethics charters, recycling content without being precise enough to transform general principles into operational compliance measures.

Moreover, under NLF regulations like the AI Act, the responsibility for developing technical standards is delegated to private bodies, the European Standardisation Organisations (ESOs), where large multinational companies are over-represented and hold significant influence. Standards developed under the NLF also carry substantial legal weight, often making them *de facto* mandatory for product compliance. In addition, these standards are typically locked behind paywalls, although the situation may evolve in the coming years after a recent case law where the Court of Justice of the European Union found that there was an "overriding public interest" in the disclosure of harmonised standards.

To finish our study of AI standards, we carried out interviews with experts from CEN-CENELEC JTC 21, who are in charge of developing the harmonised standards intended to support the AI Act. Our results show significant differences in how standardisation experts experience the drafting and negotiation of AI standards, as well as the challenges they face in addressing the demands the European Commission. Experts from large companies tend to work in teams because they have the resources to do so. On the contrary, academics and civil society experts are often frustrated in the standardisation process because they struggle to find the time and money they need to carry out their standardisation activities. The AI Act's emphasis on ethics and fundamental rights is also a source of confusion for these technical specialists, who face pressures to deliver standards on time and of good quality. As a result, the "AI Trustworthiness framework", the main standard for compliance to the AI Act, runs the risk of largely repeating repeating existing ethical and legal requirements.

The approach to European AI regulation differs from other regulations in the way it deals with risks to fundamental rights, in the way it leaves the accountability for compliance to private actors, and in how it retains a high degree of horizontality, applying AI principles to a wide variety of systems and use cases. This horizontality is an obstacle to defining clear and precise requirements. We therefore propose to refocus technical standards on vertical use cases, to allow the standards to define specific requirements that will depend on technology and context.

Résumé

L'Europe a été à l'avant-garde de l'éthique de l'intelligence artificielle (IA), en élaborant des chartes et des principes non contraignants sur l'IA « digne de confiance ». Le terme « digne de confiance » est utilisé par l'Europe pour désigner les systèmes d'IA qui sont « éthiques », « légaux » et « techniquement robustes ». L'Europe a complété ces principes non contraignants par un texte de loi sur l'IA, connu sous le nom de règlement sur l'IA, ou *AI Act*. Le règlement sur l'IA est l'un des premiers cadres légaux au monde à réglementer les systèmes d'IA dans différents secteurs et cas d'utilisation, en mettant l'accent sur la sécurité et la protection des droits fondamentaux. Pour les questions opérationnelles, le règlement sur l'IA s'appuie principalement sur des normes techniques en cours d'élaboration. L'approche européenne combine donc trois niveaux d'instruments réglementaires : les chartes éthiques de l'IA, le règlement sur l'IA et les normes techniques. Ces trois niveaux de réglementation sont censés être complémentaires, mais leur interaction est difficile. En effet, il n'est pas simple de traduire des principes éthiques en obligations légales et, enfin, en critères techniques. Dans ce manuscrit, nous posons donc la question suivante :

Qu'est-ce qui distingue l'approche européenne de la régulation de l'IA des autres réglementations européennes ?

Nous examinons tout d'abord chacun de ces mécanismes de régulation et la manière dont ils fonctionnent, indépendamment et ensemble. Tout d'abord, grâce à une analyse quantitative de 436 chartes d'éthique de l'IA, nous montrons comment certains principes d'éthique de l'IA restent spécifiques à certaines catégories de parties prenantes, tandis que d'autres ont complètement pénétré l'espace de l'éthique de l'IA. Nous mettons également en évidence les dynamiques de pouvoir sous-jacentes autour de l'éthique de l'IA, où les chartes sont utilisées comme moyen de communication et d'influence.

Le règlement sur l'IA prend racine dans certaines de ces chartes éthiques, car il a notamment été influencé par les travaux du groupe d'experts de haut niveau sur l'IA (HLEG) de la Commission européenne. Mais le règlement sur l'IA s'appuie également sur divers autres cadres européens, notamment le règlement général sur la protection des données (RGPD) et les réglementations relatives à la sécurité des produits. Les normes techniques, cependant, sont historiquement restées séparées de la législation. Elles servent plutôt d'outil économique pour façonner la dynamique du marché tout en garantissant l'interopérabilité et la qualité des produits. Depuis quelques années, elles jouent un nouveau rôle en Europe dans le cadre du nouveau cadre législatif (NLF), où les normes techniques « harmonisées » sont utilisées pour garantir la conformité avec les exigences légales. C'est l'approche que l'Europe a choisie pour le règlement sur l'IA.

Cette étude des mécanismes réglementaires montre que beaucoup de choix normatifs sont en fait effectués en bout de chaîne, dans les normes techniques. L'approche par la normalisation est fréquente dans le domaine de la sécurité des produits, mais dans le règlement sur l'IA, les normes sont également censées répondre aux préoccupations en matière de droits fondamentaux. Même lorsqu'elles définissent des termes et des propriétés apparemment techniques, les normes sont en fait chargées de valeurs. La faisabilité de l'élaboration d'exigences techniques sur des critères tels que l'équité, hérités des chartes éthiques, est donc difficile. Pour éviter de faire des choix normatifs difficiles, les organismes de normalisation jouent la carte de la sécurité en élaborant des normes qui restent à un niveau élevé. En conséquence, les normes renvoient aux chartes éthiques, recyclant le contenu sans être suffisamment précises pour transformer les principes généraux en mesures de conformité opérationnelles.

De plus, dans le cadre du NLF et du règlement sur l'IA, la responsabilité de l'élaboration des normes techniques est déléguée à des organismes privés, les organismes européens de normalisation, où les grandes entreprises multinationales sont surreprésentées et exercent une influence considérable. Les normes élaborées dans le cadre du NLF ont également un poids juridique important, ce qui les rend souvent *de facto* obligatoires pour la conformité des produits. En outre, ces normes sont généralement payantes, bien que la situation puisse évoluer dans les années à venir après une récente jurisprudence dans laquelle la Cour de justice de l'Union européenne a estimé qu'il existait un « intérêt public supérieur » à la divulgation des normes harmonisées.

Pour terminer notre étude sur les normes d'IA, nous avons mené des entretiens avec des experts du CEN-CENELEC JTC 21, qui sont chargés d'élaborer les normes harmonisées destinées à soutenir le règlement sur l'IA. Nos résultats montrent des différences significatives dans la manière dont les experts en normalisation vivent l'élaboration et la négociation des normes d'IA, ainsi que les défis auxquels ils sont confrontés pour répondre aux demandes de la Commission européenne. Les experts des grandes entreprises ont tendance à travailler en équipe car ils disposent des ressources nécessaires. Au contraire, les universitaires et les experts de la société civile sont souvent frustrés par le processus de normalisation, car ils ont du mal à trouver le temps et l'argent nécessaires pour mener à bien leurs activités de normalisation. L'accent mis par le règlement sur l'IA sur l'éthique et les droits fondamentaux est également une source de confusion pour ces spécialistes techniques, qui sont sous pression pour fournir des normes à temps et de bonne qualité. Par conséquent, le « AI Trustworthiness framework » (cadre pour l'IA de confiance), la principale norme de conformité à la loi sur l'IA, risque de répéter en grande partie les exigences éthiques et juridiques existantes.

L'approche de la réglementation européenne en matière d'IA diffère des autres réglementations par la façon dont elle traite les risques pour les droits fondamentaux, dont elle laisse la responsabilité de la conformité aux acteurs privés, dont elle conserve un degré élevé d'horizontalité, en appliquant les principes de l'IA à une grande variété de systèmes et de cas d'utilisation. Cette horizontalité est un obstacle à la définition d'exigences claires et précises. Nous proposons donc de recentrer les normes techniques sur des cas d'utilisation verticaux, afin de permettre aux normes de définir des exigences spécifiques qui dépendront de la technologie et du contexte.

Acknowledgments

I have often heard people say that a PhD thesis is a journey. It is a long travel, made of ups and downs, twists and turns, but it is also a thrilling adventure, where the PhD candidate encounters obstacles and learns to overcome them. Ultimately, it marks a turning point in a person's life. But this tale would not be complete without a mention of the people who made it possible.

The first of these people are of course my PhD supervisors, Winston Maxwell, Florence d'Alché-Buc and Tiphaine Viard, who have supported me throughout these years, giving me precious advice and guiding me as I learnt about the world of academic research. Thanks to your expertise in your respective fields, I believe we have achieved a truly interdisciplinary piece of work. I would also like to thank my jury members for their time and expertise. I am grateful to all of you for accepting to review such a particular thesis and for showing that interdisciplinarity has its place in academia. Thanks also to my doctoral committee, Raja Chatila and Valérie Beaudouin, who have followed my progress and helped me to improve my work.

I also want to thank all my teachers from the start of my graduate studies, who have helped me grow as a person, and notably my teachers at ISAE-SUPAERO and SciencesPo, who gave me a taste for all things related to AI, from its technical to its more political dimensions. I would like to thank in particular Denis Wilson – who is perhaps one of the best teacher I know – for introducing me to data science and machine learning. And thank you Denis also for allowing me to return to school on the teaching side. I have loved sharing my passion through teaching and I think that it has taught me as much (if not more) than it did the students. This teaching experience was also made possible by Véronique Steyer, who entrusted me with the management of the AI ethics and regulation seminars at École Polytechnique. Thank you also to the members of the former CNPEN, where I did my internship before this PhD. A special thanks to Catherine Tessier and Claude Kirchner for introducing me to digital ethics and for their continuous support over the years.

I am also grateful for the people I have met during this journey and with whom I had the chance to collaborate. Thank you first to my colleagues and friends in the OpAIE team, Josh, Alexandre, Astrid, Simon. I was glad to share a part of this journey with you and I wish you all the best for the future. Thank you also to Hélène, without whom I would not have been able to conduct these interviews and write the final part of this thesis. Thank you to everyone who took part in the LIMPID project, especially the people at IDEMIA. I would also like to thank the people at LNE for their warm welcome. In particular, thank you to Agnes Delaborde for making this collaboration possible. Thank you to the people I met during conferences and with whom I have developed both my network and new friendships. Thank you especially to the TALI organising team, and the two wonderful people who brought this idea to life, Chinny Sharma and Doriane Rettig. Finally, I would like to thank the JTC 21 experts who took the time to discuss with me and share their experience of standardisation. I have a feeling that our paths will cross again...

Et enfin pour finir – et dans une langue que tout le monde pourra comprendre – un grand merci à tous mes proches, à ma famille et à mes ami.e.s. Merci à mes parents et à mes grands parents pour leur soutien inconditionel. Un mot aussi pour ma Caro qui, bien que je sois la grande sœur, est aussi un modèle pour moi. Merci enfin à toutes ces personnes merveilleuses que j'ai rencontrées au cours de ma vie et que j'ai la chance d'avoir à mes côtés : à Marie, Tiphaine, Coco, Anne-So, François, Adri, Nico, toute la bande toulousaine, Romain, Marion, Jimmy, les ex-nantaises, Muriel, Alice, et toutes les autres personnes que je n'ai pas citées. Une pensée également pour celles qui ont partagées mes galères en s'embarquant elles aussi dans cette folle aventure qu'est la thèse : Camille, Alex, bon courage à vous pour la fin du voyage. Et enfin, bien sûr, un grand merci à Alice, qui m'a soutenu et conseillé tous les jours, et qui a toujours été là pour moi.

Acronyms

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
AFNOR	Agence Française de Normalisation
AHG	Ad-Hoc Group
AI	Artificial Intelligence
AI Act	Artificial Intelligence Act
AIES	AI, Ethics and Society
ANEC	European Association for the Co-ordination of Consumer Representation in Standardisation
ANFR	French National Frequencies Agency (From French: Agence Nationale des Fréquences)
ANR	National Agency for Research (From French: Agence Nationale de la Recherche)
ANSI	American National Standards Institute
Art	Article
BSI	British Standards Institution
CE (mark)	European Conformity (mark) (From French: Conformité Européenne)
CEN	European Committee for Standardisation (From French: Comité Européen de Normalisation)
CENELEC	European Committee for Electrotechnical Standardisation (From French: Comité Européen de Normalisation Électrotechnique)
CFREU	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
CNN	Convolutional Neural Network
CN IA	Commission de Normalisation Intelligence Artificielle
CNIL	French Data Protection Authority (From French: Commission Nationale de l'Informatique et des Libertés)
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CSB	Companion Standardisation Body
Dec	Decision
DIN	German Institute for Standardisation (From German: Deutsches Institut für Normung)
Dir	Directive
DKE	German Commission for Electrical, Electronic and Information Technologies (From German: Deutsche Kommission Elektrotechnik Elektronik Informationstechnik)
DL	Deep Learning
DMA	Digital Markets Act
DSA	Digital Services Act
EC	European Commission
EDPB	European Data Protection Board

EEA	European Economic Area
EFTA	European Free Trade Association
ESS	European Standardisation System
EN	European Norm
ESO	European Standardisation Organisation
ETSI	European Telecommunications Standards Institute
ETUC	European Trade Union Confederation
EU	European Union
FAccT	Fairness, Accountability, and Transparency
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPAI	General Purpose Artificial Intelligence
GPT	Generative Pre-trained Transformer
HAS (Consultant)	Harmonised standards (Consultant)
hEN	Harmonised standards (Harmonised European Norm)
HLEG AI	High-Level Expert Group on Artificial Intelligence set up by the European Commission
HLR	High Level Requirement
ICT	Information and Communications Technology
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IoT	Internet of Things
ISO	International Organization for Standardization
IS	International Standard
IT	Information Technology
ITU	International Telecommunication Union
JRC	Joint Research Centre (of the European Commission)
JTC	Joint Technical Committee
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LSTM	Long Short-Term Memory
MDCG	Medical Device Coordination Group
ML	Machine Learning
MSP	European Multi-Stakeholder Platform on ICT Standardisation
NAB	National Accreditation Body
NEN	Royal Netherlands Standardization Institute (From Dutch: Stichting Koninklijk Nederlands Normalisatie Instituut)
NGO	Non-Governmental Organisation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NSAI	National Standards Authority of Ireland
NSB	National Standards Body
OECD	Organisation for Economic Co-operation and Development
OJEU	Official Journal of the European Union
OOXML	Office Open Extensible Markup Language
PIP	Poly Implant Prothèse (from French)
Reg	Regulation
RMF	Risk Management Framework

RNN	Recurrent Neural Network
SAR	Specific Absorption Rate
SC	Sub-Committee
SCC	Standards Council of Canada
SDO	Standards Developing Organisation
SME	Small and Medium-sized Enterprise
SR	Standardisation Request
SRAHG	Standardisation Requests Ad-hoc Group
StandICT.eu	ICT Standardisation Observatory and Support Facility in Europe
TC	Technical Committee
TEU	Treaty on European Union
TFEU	Treaty on the Functioning of the European Union
TG	Task Group
TR	Technical Report
TS	Technical Specification
UK	United Kingdom
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNI	Italian Standards Body (From Italian: Ente Italiano di Normazione)
US	United States
VDE	Association for Electrical, Electronic and Information Technologies (From German: Verband der Elektrotechnik, Elektronik und Informationstechnik)
W3C	World Wide Web Consortium
WEF	World Economic Forum
WG	Working Group
WTO	World Trade Organisation

Contents

INTRODUCTION

II A FOCUS ON STANDARDS TO REVEAL TENSIONS BETWEEN REGULATORY MECHANISMS

DISCUSSION AND CONCLUSION

List of Figures

List of Tables

INTRODUCTION

Chapter 1
Introduction to the regulation of AI: definitions and thesis scope

Artificial Intelligence (AI) has been part of our daily life for some years now. Browsing the web, spending time on social media, watching a movie on a video streaming platform, finding the best route on a GPS, even unlocking a phone with fingerprint or a face... All these are computer science applications often supported by the use of an AI. In computer vision, image classification and biometric recognition are already well-established technologies, with promising new applications, particularly in healthcare.

The term "AI" has always been fuzzy and regroup changing realities. However, nowadays, when the general public or the media talk about AI, it is mostly about Generative AI (GenAI). In November 2022, the American company OpenAI[1] took the world by storm by releasing its Large Language Model (LLM), accessible to all on an online platform: ChatGPT[2]. The chatbot[3] is capable of answering any question with a form of language close to that of humans, which immediately fired imaginations[4].

But while ChatGPT set the trend for GenAI, companies have not stopped there. After the release of subsequent language models such as Gemini or Copilot, or image generation models such as Dall-e, Stable Diffusion or Midjourney, the internet is now getting to grips with AI, using it to generate music and videos. AI, which a few years ago was just a scientific field, is now very popular with the general public. This hype led to an increase in investment by public and private bodies: the AI race had begun.

Nevertheless, civil society organisations and researchers alike had been sounding the alarm for quite some time, stressing the potential dangers of these technologies (Wiener, 1950; Good, 1966; Moravec, 1990; Russell, 2019). In April 2021, these calls were answered when the European Commission published draft legislation to regulate AI systems: what was to become known as the "AI Act" (European Commission, 2021e). The race to AI had transformed into a race to AI regulation (Smuha, 2021b).

In this thesis, we discuss various means of regulating AI, which span on three domains: the ethical sphere – with AI ethics charters, the legal sphere – with the EU AI Act and the technical sphere – with technical standards. We show that, although initially intended to be separate and from the most general to the most specific, the lines between these areas are becoming increasingly blurred, with the charters, law and standards becoming more and more alike. In Europe in particular, the adop-

[1] `https://openai.com/`

[2] `https://openai.com/chatgpt/`

[3] A chatbot is a software designed to interact with a human by imitating a conversation through text or voice.

[4] See Section 1.1.2.

tion of the AI Act has fuelled this trend, proposing a binding legal act that remains true to its ethical roots while attempting to establish a technical compliance framework. But the AI Act is not the only means of regulation, especially as Europe is insisting on diversifying frameworks and creating new forms of regulation, outside of mandatory law. Harmonised standards (hEN) are the perfect example of this European mode of regulation: a regulation which is not strictly the work of state or supra-state actors, but also that of private actors.

Harmonised standards are not new to EU regulation; they have been used for decades in the field of product safety, alongside the European Conformity (CE) mark, which certifies compliance with EU legislation and allows products to be distributed on the European market. Yet, with the AI Act, there is a trend towards addressing ethical, social and fundamental rights issues through standards. This raises the question of what standards can actually contain and whether current standardisation stakeholders are equipped and legitimate to deal with these issues.

1.1 *Context and background*

In this section, we present the main concepts of this thesis: "artificial intelligence" and "regulation". We show that these two concepts are highly polysemous[5] and we outline the scope that we chose to consider for this thesis. We first highlight the need to regulate AI, as these new technologies raise both technical challenges and dangers for individuals and society as a whole. But we also analyse the difficulties encountered in regulating AI, which we attribute mainly to the the complexity of AI technologies, and the diversity of regulatory instruments available. To regulate AI efficiently, it is necessary to understand the relationship between all these regulatory instruments.

1.1.1 *A short history of AI*

If the term "artificial intelligence" was coined less than a century ago, the idea on which AI is based is thousands of years old. In ancient Greece, the automata of the god Hephaistos can be seen as the ancestors of robots, such as his golden servants, created to assist him, or the bronze giant Talos, guardian of the island of Crete (Marcinkowski and Wilgaux, 2004). This idea of divine creation is also present in monotheistic religions, such as the myth of the Golem in the Bible. Throughout the 19th century, literature continued to spread the idea that beings could be born artificially, as in Pinocchio (Collodi, 1881) and Frankenstein (Shelley, 1993). Science fiction in the cinema followed the same path in the 20th century, erecting superpowered robots taking control by force and acting as a weapon of destruction for humanity, such as HAL in 2001: A Space Odyssey (Kubrick, 1968), or in the Terminator movies (Cameron, 1984). While our imaginations continue to influence our vision of AI, they are a far cry from modern technologies.

In 1950, in a famous article entitled "Computing machinery and intelligence", published in the journal Mind, Alan Turing proposed to consider the question: *"Can machines think?"* (Turing, 1950). In the though-provoking Turing test, or imitation game, the mathematician proposes an experiment in which the machine has to deceive a person, making him or her believe that he or she is talking with another human being. This article is considered to be the first step towards modern artificial intelligence, even though the term was not coined until later. In 1956, the word

[5] For the different meanings of "artificial intelligence", see Cardon et al. (2018) and for the different meanings of "regulation", see Morgan and Yeung (2007).

first appeared at the Dartmouth Workshop on AI, a brainstorming event attended by famous mathematicians such as John McCarthy, Marvin Minsky and Claude Shannon, among others (Dartmouth, n.d.). From then on, AI was considered as an emerging field, distinct from the broader field of cybernetics led by Norbert Wiener in the 40s. But at that time, AI was far from having reached the potential it has today. In 1956, AI is defined as: *"the construction of computer programs that engage in tasks that are currently more satisfactorily performed by human beings because they require high-level mental processes such as: perceptual learning, memory organization and critical reasoning"* (Bertrand, 2024).

The subsequent history of AI has been one of back and forth, oscillating between technological breakthroughs and disappointing periods of stagnation: the AI winters (Cardon et al., 2018). However, these different periods are also the result of a constant normative redefinition of the concept of AI. The ferment of ideas that emerged in the 40s and 50s gave rise to two approaches: symbolic and connectionist AI. The connectionist movement is based on an invention by Warren McCulloch and Walter Pitts in 1943: the perceptron, an artificial neuron later used as the basis for neural networks. However, in 1969, the limitations of the perceptron led to the first AI winter. It is only in 1989, that Yann LeCun would revive the movement with the invention of backpropagation (LeCun et al., 1989), a necessary mechanism for the learning of neural networks. But resources remain limited and neural networks were not as performant as their symbolic counterparts. Indeed, the late 70s and 80s had seen the emergence of expert systems, computer systems designed to solve complex problems using predefined rules or a body of knowledge, and considered a cornerstone of the symbolic movement. For long, they remained the state-of-the-art of the AI field, up until the 2010s.

After the explosion of large-scale data in the 2000s, the ground was set for the AI community to start taking part in international challenges. Research teams would compete to achieve higher accuracy on a benchmark of given data, such as the ImageNet challenge in 2010 on several visual recognition tasks. In 2012, AlexNet, a convolutional neural network (CNN), took the world of artificial intelligence by storm by winning the ImageNet challenge by a wide margin over the benchmark expert systems. The hegemony of machine learning, and in particular deep learning, was only just beginning, and neural networks would subsequently be used for a large number of tasks. In particular, remarkable advancements in natural language processing (NLP), with the publication in 2017 of the transformer model, are the basis of current AI models, such as ChatGPT. These recent advances in AI have been made possible by the interweaving of academic and industrial research, which has made the ecosystem more complex and given rise to conflicts of interest.

1.1.2 *A criticised term*

The imaginary worlds created by science fiction have had a lasting effect on the general public's representation of what AI is, or what it could be. The possibility of an AI capable of matching or even surpassing human intelligence is not new (Wiener, 1950; Good, 1966; Moravec, 1990). It is referred to by Ganascia (2017) as the *"myth of singularity"*. AI systems with human-like intelligence have been given a variety of names, from "strong AI" (Searle, 1980) to "artificial general intelligence" (AGI) (Goertzel and Pennachin, 2007) to "superintelligence" (Bostrom, 2014). These movements, whose main players belong to American transhumanist institutes[6], base their arguments more often on storytelling than on scientific facts (Gebru and Torres, 2024).

The rise of GenAI in recent years has only fuelled this myth because of its perceived ability to perfectly mimic the human ability to write, paint or create in general. As well as misleading the general public, this myth also affects specialists in the field, such as well known computer scientists in BigTech companies, who have ensured the public that AGI is almost, if not already, here (Maruf, 2022).

But GenAI is based on probability, not human reasoning. If a sentence written by an LLM seems to have been created by a human, it is only because the system calculates the most likely word that should follow, but the substance of the text could be completely meaningless. The best proof of this is the difficulty of ChatGPT to do maths or pass basic logic tests (Lohr, 2024), an issue that is not new since it is shared by all text generators (Weizenbaum, 1966).

Some experts have argued that AI is *"neither artificial nor intelligent"* (Crawford, 2021, Introduction), calling for a different name to be used. Indeed, the term artificial intelligence has been criticised for being too anthropomorphic (Tessier, 2021)[7] and misleading people into thinking that these technologies can have a resemblance with human intelligence.

The first problem that anthropomorphisation can pose is public deception and misinformation, as "hallucinations" caused by chatbots could be accepted as is, without questioning the veracity of the information provided (Metz, 2023). Furthermore, the media hype surrounding AGI nourishes fear among part of the population. But most importantly, these debates occlude the real dangers of AI: that systems can fail and cause a variety of harms[8], i.e. technical systems embed social values (Birhane et al., 2022).

Most of these issues are due to a lack of public education about how current AI systems work. However, there is a risk that anthropomorphisation will further confuse part of the population. In particular, the comparison with intelligence does not help non-experts to understand the intrinsic limitations of these technologies and the differences with our own reasoning abilities.

Although criticised, the term AI has now been widely adopted by the general public and experts alike. In particular, it is the accepted term for these technologies in the fields of ethics, law and standardisation. Around the world, guidelines for these technologies are called "AI ethical charters". European law now has an "AI Act" (European Parliament and Council, 2024b) and other countries are likely to follow[9]. To oversee the implementation of the AI Act, Europe now officially has an "AI Office" and an "AI Board"[10]. Even in the field of standardisation, dominated by technical experts, both at national and European or international level, committees working on these technologies use the term "AI"[11]. The trend is almost impossible to resist.

Because of the now official nature of this nomenclature, we have chosen in this thesis to retain the term "artificial intelligence" to designate these technologies, despite the criticisms. The rest of this introduction will therefore define exactly what we mean by AI: what technologies we consider and how they work.

[6] To refer to this line of thought, Gebru and Torres (2024) coined the term "TESCREAL": *"transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism"*.

[7] Anthropomorphic means *"described or thought of as having a human form or human attributes"* (Merriam-Webster Dictionary, n.d.a).

[8] See Section 1.1.6

[9] See Chapter 9, Section 9.7.3.

[10] See Chapter 3, Section 3.5.4.

[11] For instance ISO/IEC SC 42 or CEN-CENELEC JTC 21. For more information on the different standardisation groups working on AI, see Chapter 5, Section 5.3.

1.1.3 *AI in computer science*

Related terms

Before we define AI, let us talk about some related terms which need to be understood in order to understand AI itself.

Modern AI is mostly based on machine learning (ML), models that need data to function. "Data" is broadly understood as *"information in digital form that can be transmitted or processed"* (Merriam-Webster Dictionary, n.d.f)[12]. Data can be scalars – such as temperature or market prices; it can be vectors – such as forces or trajectories; it can be images, audio files, categories, and so on.

Another important term is "algorithm". Although AI systems can be seen as a family of algorithms, many algorithms have nothing to do with AI. Algorithms is usually considered as a sequence of simple, non-ambiguous steps to solve, exactly or approximately, a problem (IGI Gloabl, n.d.).

We often hear that the difference between traditional algorithms and AI is precisely AI's ability not to be explicitly coded to do something.

But in AI models, the steps are explicitly coded, however, there is paradigm shift: whereas non-AI algorithms are deductive machines – where the input and the computer program produce outputs, AI algorithms based on machine learning[13] are inductive machines, where the input data and an externally decided objective produce the computer program (Cardon et al., 2018). This objective may be, for example, to maximise prediction accuracy, or succeed at a certain task. In the case of machine learning models, this objective can be carried, for example, by the loss function in supervised learning, or by the reward function in reinforcement learning[14]. These functions, and therefore the objective of the model, is encoded explicitly by humans.

Symbolic AI v. Machine Learning

One of the greatest difficulties when working on AI is actually defining what AI is in the first place. There are almost as many definitions as there are people, but general ideas tend to emerge, notably amongst experts who share the same field of expertise. One of these many definitions is that AI is a research field, a branch of computer science (Merriam-Webster Dictionary, n.d.b).

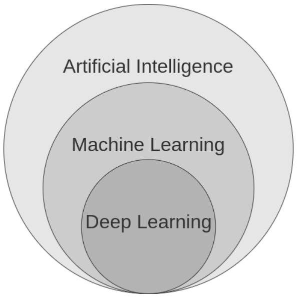


Figure 1.1: Widespread representation of the different types of AI. However, the diagram is a rather inaccurate simplification of the field.

From there, everyone uses their own diagram to show the different sub-domains of AI. One of the most widely circulated diagrams is an image with a circle representing AI, another circle inside for machine learning and another inside for deep learning, as shown in Figure 1.1. However, this is a simplified version of reality. Therefore, we have chosen the following representation, shown in Figure 1.2[15].

AI is often separated between Symbolic AI, also called Good Old Fashioned AI (GOFAI), and Machine Learning (ML). These two branches of AI are largely derived from the two main approaches that emerged in the 1950s: the symbolic approach and the connectionist approach. Today, however, "AI" is mostly used as a synonym for machine learning[16].

One way of differentiating Symbolic AI from ML is to think about the logical reasoning behind them. Symbolic AI is based on deductive reasoning, where general rules or a body of knowledge are used to reach a conclusion about specific examples. On the contrary, ML is based on induction. In inductive reasoning,

[12] By comparison, in law, data is defined in the Data Market Act as *"any digital representation or compilation of acts, facts or information"* (European Parliament and Council, 2022).

[13] See next paragraph for the difference between Symbolic AI v. Machine Learning

[14] See below for a summary on learning mechanisms.

[15] This figure is inspired by a course I took during my engineering studies. I therefore would like to thank Mr Nicolas Drougard and his course "Introduction to Machine Learning" at ISAE-SUPAERO.

[16] It should be noted, however, that many systems that make the headlines under the name of "AI" are not in fact AI (Knight, 2019a). One example is the COMPAS algorithm (Angwin et al., 2016). Similarly, today many companies that claim to be doing AI are not actually using AI, but are simply using the term for marketing purposes.

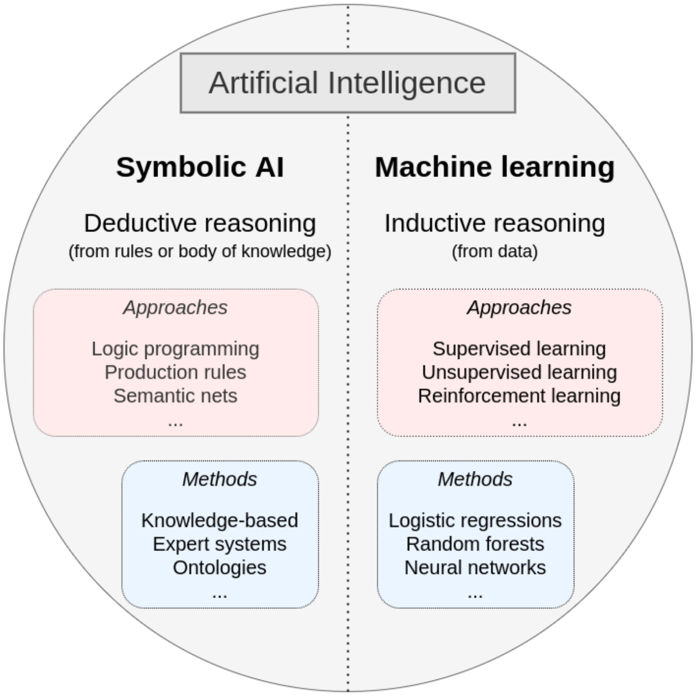


Figure 1.2: New representation of the different types of AI.

specific examples – here, the data – are used to deduce general rules. The aim of machine learning models is therefore to find patterns and correlations in the data in order to generalise and find a solution that works in most cases: ML models are probabilistic.

As we have seen, one way of approaching the definition of AI is therefore to draw up a list of everything that can be considered AI, including symbolic approaches and machine learning approaches.

However, as the rest of this thesis focuses primarily on AI regulation and the AI Act, we will leave the definition of AI to the legal domain. However, we will first look at how ML models work in order to better understand the main technology behind AI, as well as its shortcomings and the dangers that come with it.

Learning mechanisms

Figure 1.2 also makes a distinction between the categories of approaches and the methods. The approaches are the general paradigms, while the methods are the techniques used to implement them. In ML for instance, there is a distinction between learning mechanisms and algorithmic models.

There are three main types of learning mechanisms. To begin with, supervised learning models are trained on labelled data and their objective is to predict the label[17] of a new data point. They are used for (i) classification tasks, where the target variable[18] is a categorical variable, such as predicting whether or not a person will repay their loan; and (ii) regression tasks, where the target variable is a continuous variable, such as the prediction of market prices for products. In supervised learning, the model learns to make the best predictions during training using a loss function which provides the model with feedback and updates the model parameters to improve predictions[19]. Prediction on a new data point is made at a later stage, once the model has been fully trained[20]. This new data point on which the

[17] Labels are the possible values of the target variable. They are also called "ground truth" and are created by humans.

[18] To perform supervised machine learning tasks, data must be associated with variables, either feature variables, which are used to make the prediction, or target variables, which are the outcome the ML system is trying to predict. Each data point is thus associated with several feature variables and one target variable. For instance, when a bank wants to predict if a person (the data point) is likely to repay their loan (the target variable), the bank will look at the person's level of income, their age, their marital status, etc (the feature variables).

[19] In neural networks, this mechanism is called "backpropagation".

[20] With the exception of systems that use continuous learning to continue to learn after deployment.

prediction is made must not be part of the training set.

The second main type of learning mechanisms is unsupervised learning. In unsupervised learning, the data provided to the system is not labelled. The objective of the model is therefore to find coherent structure in the data, for example by grouping it into different clusters[21], detecting anomalies, or compressing data.

Finally, the third main type of learning mechanism is reinforcement learning, where an agent learns to best behave in an environment. Learning is based on trials and errors. When the agent performs a good action, it receives a reward, via a reward function. It then learns the optimal policy for optimising this reward. These systems are best known for their use in board games[22] and video games, but have also led to major advances in healthcare[23].

However, not everything can appear on Figure 1.2, as we cannot pretend to make an exhaustive list of all learning mechanisms and ML methods. Obviously the frontier between the two domains of symbolic AI and ML is less strict than on this representation, as some systems uses hybrid approaches, a mix of the two worlds. In addition, it should be noted that there are not exclusively three types of learning, with some experts considering as distinct learning mechanisms those that mix different approaches or that are sub-parts of the three main approaches. This includes for instance semi-supervised learning, for mechanisms which fall between supervised and unsupervised; transfer learning, where a dataset from one domain is used to help make predictions in another domain; self-supervised learning where the model learns one part of the input from another part of the input; few-shot (respectively one-shot) learning where the model learns by training on a very small number of (respectively one) example(s).

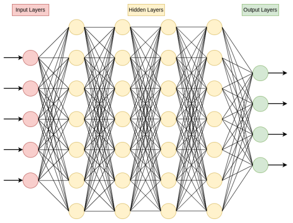


Figure 1.3: Representation of a fully-connected neural network. Each neuron from a given layer is connected to all the other neurons from the previous layer. This is a simple structure and most neural networks today use more varied and complex structures.

Deep Learning, Generative AI and Foundation models

Another category that does not directly appear on Figure 1.2 is deep learning. Indeed, deep learning is a learning mechanism that can be both supervised and reinforcement-based, and which specifically uses neural networks with a large number of layers, i.e. deep neural networks, as the one represented in Figure 1.3. Each artificial neuron performs a very simple operation: the weighted sum of all neurons values before it, such as represented in Figure 1.4.

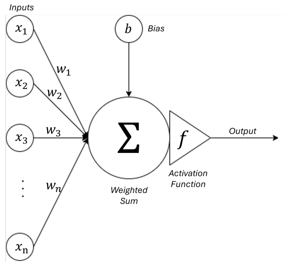


Figure 1.4: Representation of an artificial neuron. The neuron fired when the weighted sum of all previously connected neurons is above a certain threshold, calculated by the activation function.

Once the value of the neuron is calculated, it goes through an activation function which determines if this value is enough to fire the neuron. The combination of all operations performed by all the neurons in the network allows to recreate almost any complex phenomenon: this is known as the universal approximation theorem[24]. This property makes neural networks easy to use for a wide variety of tasks, on images, text or other data types[25].

Finally, GenAI does not appear either on Figure 1.2, as it falls into existing categories of learning[26]. The precise nature of GenAI's current models is unknown to the general public because they are subject to industrial secrecy. However, some of methods which derive primarily from neural networks can be used to generate outputs. For example, GANs can be used for image generation and transformers, such as the GPT[27] family, for text generation.

These models, specifically in the field of Natural Language Processing (NLP) for text generation, are mostly trained through self-supervised learning. This can take the form of a masked language modelling task where the model is trained to predict

[21] This ML task is called "clustering".

[22] See AlphaGo (DeepMind, n.d.b).

[23] See AlphaFold (DeepMind, n.d.a).

[24] The universal approximation theorem (Hornik et al., 1989) states that a neural network with a sufficient number of neurons can approximate any continuous function.

[25] We can cite for instance, Convolutional Neural Networks (CNNs) for image classification; Generative Adversarial Networks (GANs) for image generation; Recurrent Neural Networks (RNNs) for speech recognition; Long Short-Term Memory (LSTM), as type of RNNs, for image captionning; Autoencoders for dimensionality reduction; Transformers, for text generation, etc.

a missing word in a sentence based on the surrounding context[28].

Some models can use a wide variety of data input types and use them for different downstream tasks. They are known in computer science as *"foundation models"* (Bommasani et al., 2021). From a technical standpoint, foundation models are enabled by transfer learning: the model is pre-trained on a surrogate task, then adapted to the downstream task via fine-tuning. Fine-tuning involves re-training the model on new data specific to the downstreamed task. But what makes foundation models powerful today is their scale: they have huge computing capacities, efficient model architecture and large amounts of training data.

1.1.4 AI in law

AI systems

Unlike in computer science, there is only one official way of defining AI in the legal sphere: by using the definitions in the laws in force. In European law, the AI Act (European Parliament and Council, 2024b) gives a definition of an *"AI system"* (art 3(1) AI Act):

> ***Definition***
>
> ***AI system (AI Act).*** *A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*

The definition uses fairly technical terms such as *"inputs"* and *"outputs"*, but what seems to separate a simple algorithmic system from an AI system under the AI Act is the notion of *"autonomy"*[29]. Although we have no definition of this term in law, in a general sense it can mean: *"the quality or state of being self-governing"* (Merriam-Webster Dictionary, n.d.c).

With this definition alone, it is not clear whether symbolic AI models are included or just machine learning models. Indeed, even if the objectives can be *"implicit"* or *"explicit"*, the definition insists on the need for inference. This is further exacerbated by Recital 12, which states that the definition of AI systems *"should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations"* and that *"a key characteristic of AI systems is their capability to infer"*. However, the same recital goes on to state that: *"The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved"*. Therefore, both Symbolic AI and ML methods should be covered by the AI Act[30].

Systems or models?

The AI Act makes a point of regulating specifically AI *"systems"*, not AI *"models"*, with the exception of General Purpose AI models[31]. The AI Act explains the difference between AI systems and AI models, as Recital 97 states that: *"Although AI models are essential components of AI systems, they do not constitute AI systems on*

[26] Mostly, unsupervised, self-supervised, or transfer learning.

[27] GPT stands for Generative Pre-trained Transformer.

[28] This is how the language model BERT is trained (Devlin et al., 2019)

[29] This definition has evolved greatly since the first proposal of the Commission. See Chapter 3, Section 3.6.1.

[30] However, this conclusion will have to be verified in the future, when we have further case law on which to rely.

[31] See next paragraph.

their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. AI models are typically integrated into and form part of AI systems". Recital 101 further emphasises that models can be integrated into products, but do not constitute products in themselves. This is why models are only mentioned when training is involved (Recitals 67 and 76). *"Systems"* are the end products, while *"models"* enable the systems to function. The main reason for this distinction is that the AI Act aims to regulate only products that are put on the market, so they must be final end products and not components of such products.

General Purpose AI and Generative AI

The AI Act makes a distinction between traditional AI systems and *"General Purpose AI"* (GPAI) systems. GPAI systems are defined as: *"an AI system [...] which has the capability to serve a variety of purposes"*.

GPAI systems are notably based on GPAI models, defined in Article 3(63) as:

> ***Definition***
>
> **GPAI model (AI Act).** *An AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications.*

The definition insists on the importance of the wide variety of tasks such a system can solve and the wide range of applications to which it can be applied. We can also note a desire to describe in more detail how GPAI models can work, specifying that this applies to systems working on large amounts of data and using self-supervised learning[32].

In particular, the notion of GPAI in the AI Act corresponds to the computer science concept of *"foundation model"*[33], a name which was briefly used some amendments to the AI Act[34], before being removed from the final version.

Finally, while GenAI does not have a specific definition, *"AI systems generating synthetic audio, image, video or text content"* are cited in the text (Art. 50 AI Act), and should not be confused with GPAI systems, a category of systems to which they may or may not belong. Indeed, while most GenAI systems today are capable of processing different types of data, we could imagine a system which only generates a certain type of data.

1.1.5 Technical shortcomings of machine learning

ML systems have technical limitations that we need to be aware of in order to understand the growing problems associated with their use[35].

The need for "good" data

ML models are very sensitive to input data and they need large amounts of data to train. For a long time, computer scientists though they could improve the accuracy of their AI models by simply using bigger models. However, these bigger and more complex models required more data to train: this is known as the curse of dimensionality (Bellman, 1961).

[32] For a technical definition of those terms, see Section 1.1.3.

[33] To understand how foundation models work, see Section 1.1.3.

[34] The term *"foundation model"* was introduced by the European Parliament during the negotiations on the IA Act and appears in various amendments to the text. A definition is notably given in amendment 168 (European Parliament, 2023a).

[35] For more information on the technical limitations of machine learning, see:

> Gornet, M. Limitations techniques du machine learning et choix de conception. In Nevejans, N. *Décision humaine, décision de l'IA*, to be published.

To improve the performance of AI models, quantity is not enough: data quality is also of crucial importance. The correct operation of AI models is based on statistical hypotheses that are never actually true in practice. This is the case, for example, of independent and identically distributed data. Indeed, AI models need a variety of input data, if possible "representative" of real-world data[36], otherwise biases may result from data sampling. It is also often desirable to have balanced data between the different classes[37]. If there is a gap in the training data, the model will misbehave in the area where the data is missing. In general, if the training data is not "good", for example if it is not balanced or representative, the model will not make good predictions either.

The lack of data can lead to "bias", particularly when a system is used on different groups of people. However, the term bias is highly polysemous and we should separate biases which are necessary for the system to function, and biases which are undesirable and can cause harm[38]. Here, the word "bias" corresponds to *"a systemic error"* (Merriam-Webster Dictionary, n.d.d), or a *"systemic difference in treatment of certain objects, people, or groups in comparison to others"* (ISO/IEC, 2021). But lack of data alone is not responsible for bias, as some studies have shown that ML systems can be biased even when they are trained on perfectly balanced datasets (Albiero et al., 2020; Gwilliam et al., 2021). Sources of bias can also be found, for instance, in the way the model is trained or deployed (Suresh and Guttag, 2021).

In addition to raw data, labels are necessary for supervised learning. Indeed, the aim of the model is to predict a label as accurately as possible. Learning can therefore be, at most, as effective as label accuracy. Labels are, however, highly subjective. For instance, for a classification task, both the label categories and the assignment of a label to a data instance are human choices[39].

Additionally, more and more models are relying on human feedback to improve their performance. This is the case, for instance, with many large language models, such as those of OpenAI, where techniques such as Reinforcement Learning from Human Feedback (RLHF) are used to fine-tune Generative Pre-trained Transformer (GPT) models (Ouyang et al., 2022). AI is therefore not totally artificial.

Probabilistic models

However, shortcomings do not always come from the data. Indeed, all supervised ML models seek to find correlations in the data and learn the rules that will optimise the accuracy of the results. But optimising the model is a difficult task. On the one hand, if the model learns too much on the same training data, it will be unable to generalise to new data – this is known as overfitting. On the other hand, if it does not learn enough, the model will be under-optimised for the task in hand – this is known as underfitting. Therefore, developers of ML models need to strike the right balance between overfitting and underfitting. However, as the aim of the training is to be able to generalise, and due to their probabilistic nature, errors are inherent to ML models. This is particularly true for cases that fall outside the norm. This is because the fewer examples of similar cases the system has seen in the data, the less likely it is to be able to predict correctly. This can cause problems for people who fall outside the norm, on whom these systems might not work properly[40]. This calls into question the very use of these technologies, which are incapable of working for the entire population.

Because they are based on the correlation of data, ML models are not able to

[36] A dataset is "representative" of another if it is smaller in size and capture most of the information from the original dataset (Borovicka et al., 2012). In other words, the distribution of the two datasets must be close.

[37] Other solutions involve using models that work on unbalanced data or rebalancing the data.

[38] See for instance, the case of the Amazon hiring algorithm presented in Section 1.1.6.

[39] For instance, Kozyrkov (2024) takes the example of a classification task between "cat" and "non-cat" in which the image of a tiger is presented to the annotator. Two human choices have been made: choosing the two restrictive categories without considering a third option and, after accepting these categories, choosing whether the tiger can indeed be considered a cat.

[40] The LGBTQ+ community is a case in point. For more information, see:

> Gornet, M., Viard, T. (2023b). Queer identities and Machine Learning. `https://hal.science/hal-04763352`

invent, they simply create rules from previous data. One of the main limitations of ML approaches is therefore that they are based on the postulate that the past is equal to the future. This makes AI virtually useless for predicting very rare events[41]. These technologies need constant re-training and the use of new data in order to evolve over time. And even in this case, the biases present in the previous data can persist, even when new data is added.

Opaque models

The general rules ML systems can learn are represented, for instance, by the branches of a decision tree, or the weights of a linear regression. However, these learned rules are not always – or even, very rarely in the case of complex models – accessible, at least not in a comprehensible form. In neural networks for instance, these rules are the weights and biases of each individual neuron, which do not explain in human-understandable terms how the system arrives at a certain output. This makes complex models very opaque, which earned them the name "black box", as we know what goes in and out but not what is happening inside. The ML research community has therefore focused on developing "explainability methods" to make complex models more comprehensible by humans [42].

The opacity of ML models means that humans, even the developers of these models, do not necessarily understand the correlations established by the models, so that "bad" rules can be learned without any possibility to verify them. An example of a model which had learnt "bad" rules was discovered by (Ribeiro et al., 2016) in a famous paper on the explainability of ML. During an image classification task, Ribeiro et al. (2016) tried to understand why the model misclassified an image of a husky as a wolf. Using their explanation method, they highlighted the pixels that played the most important role in the classification and realised that the snow in the background had a significant influence on the decision. Indeed, in the training data, most wolves were found on snow, so the model learned to associate these elements, even though the fact that the animal was on the snow did not mean that it was a wolf[43].

However, when the model works well, there is often no reason to question the rules it has learned. As a result, some models may work apparently well, but for the wrong reasons. This phenomenon is known thanks to Clever Hans, a horse reputed to be "intelligent" because it could apparently make calculations by hitting the ground. However, it was discovered later that the horse was simply reading its master's face when it had to stop (Pfungst, 1911), showing empathy in reading non-verbal cues, rather than reasoning. Similarly, Google Flu, a tool supposed to predict flu epidemics based on keywords typed into Google Search by web users around the world, was discontinued by Google a few years after its release (Lazer et al., 2014). The algorithmic model had several problems, such as its incapacity to dynamically adapt to new web search trends, or the presence of spurious correlations, such as between flu predictions and basketball season[44].

1.1.6 The social dangers of AI

While the technical shortcomings of AI might be seen as trivialities, combined with the misuse of this technology, they can create high risks of danger for individuals and society as a whole. Indeed, over the years, concerns have arisen about the

[41] Events that Taleb (2010) describes as "black swans", such as 9/11 or, more recently, the Covid-19 pandemic.

[42] However, this approach is criticised by some researchers, such as Rudin (2019) who argues that ML developers should focus on using simpler models that are intrinsically explainable, rather than using black-box models. Indeed, current explanation methods are very limited and can be misunderstood.

[43] This example shows that correlation is different than causality: the presence of snow and the wolf prediction are correlated, but having snow does not cause the animal to be a wolf.

[44] The start of basketball season in the United States generally occurs at the end of October, when the first cases of flu appear. The two phenomena are therefore correlated, even if there is no causality. Google Flu has therefore learned to predict an increase in the number of flu cases when web users search for terms relating to basketball.

harms that AI systems can cause, with breaches of fundamental rights and safety issues have sparked outrage from the general public and the media. In this section, we provide a short, non-exhaustive list of the dangers and challenges posed by AI systems, in order to understand why regulation is necessary.

Biases and the risk of discrimination

One of the most common examples of damage caused by AI systems is discriminatory bias. For example, facial recognition algorithms have been found to be less accurate for people of colour (Buolamwini and Gebru, 2018)[45] or queer individuals (Scheuerman et al., 2019), on several categories of tasks, such as detection, authentication and gender classification. This is in part due to the fact that ML models are generally trained on datasets that are not balanced in terms of gender or race[46]. These inaccurate results can cause damage which goes beyond the simple failure of the system.

In 2021, Robert William, an African-American man with no criminal record, was arrested at his home in front of his family for allegedly stealing watches. The shop was using an AI-based video surveillance system which mistook Robert William for a wanted criminal (Hill, 2020b). A series of similar "errors" continued to be reported across the United States for several years, with the falsely accused systematically being African-Americans (Anderson, 2020; Hill, 2020a). Five years later, in June 2024, Robert William finally got closure in a trial which costed the city of Detroit $300,000 (Sahouri, 2024).

However, biases alone do not always cause discrimination, this is also due to the use we make of this technology. In the case of Robert William, the failure was caused both by the algorithm not recognising the right person and by the police not taking the time to check the information.

Beyond facial recognition, image classification also suffers from serious prejudice against the African-American population, as shown by the Google Photo scandals. The system mistakenly classified black men as gorillas (BBC News, 2015), or, when shown photos of hands holding thermometers, classified them as firearms if the hand was that of a black person (Kayser-Bril, 2020). Although harmful, these errors do not directly threaten people's lives. However, this could be the case for healthcare applications. For example, skin colour could have an impact on the accuracy of AI-based systems in detecting skin cancer (Rezk et al., 2022).

Sometimes, it is not the lack of data which causes biases to appear but the correlation between a decision variable and a protected attribute[47]. For example, Amazon was widely criticised in the media for using a recruitment algorithm that penalised female applicants (Dastin, 2018). The American company used data from former employees – mainly men – to train its system and determine the characteristics of a "good" employee. The algorithm subsequently learned that being male, a criterion statistically correlated with employability in the company, was a "good" decision criterion for job applicants, even though gender was not explicitly one of the criteria used for the decision.

But the Amazon's decision system is not the only one to exhibit biases and trigger discussions about what constitutes a "good" decision. In 2019, Apple was also under scrutiny for its credit card system that offered smaller credit lines to women than men (Knight, 2019b). The case is now closed, with the New York State Department of Financial Services investigation concluding that there was no gender

[45] And even more so for women of color.

[46] Here we use the term "race", which is mostly used in the Anglo-Saxon world, to designate groups of people who share physical traits due to a common ancestry. The term "ethnicity" may also be used, although the latter term is broader and also includes elements such as national, tribal, religious, linguistic or cultural origin or background (Merriam-Webster Dictionary, n.d.h).

[47] Protected attributes are features that should not be used by the model for decisions. They are often features protected by law, such as skin color, gender, sexual orientation, religion, disability, and so on.

discrimination. However, as the credit decisions were sometimes justified by elements of the credit history, the case still raises questions about what can be used as a legitimate variable (John-Mathews, 2021). The same question can be asked of car insurance companies, which have been accused of raising their rates in minority neighbourhoods. While insurance sellers claim that rates increase with the risk of an accident, the price disparities between minority and white neighbourhoods appear to be greater than the differences in risk. The use of a biased algorithm by sellers could be at the root of this unequal treatment (Angwin et al., 2017).

Where biases are most dangerous, however, is when they contribute to feeding an entire real-world system which is already biased. The best-known example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) assessment, an algorithmic system used in the United States to assess the risk of criminals reoffending and to help judges take the appropriate sanctions. However, some studies reveal that the system was biased against African-Americans, who received higher risk scores (Angwin et al., 2016). While these results have been criticised[48], they remain at the centre of a wide-ranging debate on the reinforcement of inequalities, with black people being statistically incarcerated more often as a result of a system which still contains systemic discrimination (Alexander, 2012). The criteria taken into account by the algorithm, even if race does not appear, include the number of times a person has been arrested or family members already in prison, which are strongly correlated with blackness. As such, they are likely to contribute to higher risk scores for black people, reinforcing the loop.

The same problems arise in recruitment, where gender-neutral advertisements promoting job opportunities in science, technology, engineering and mathematics are systematically shown more often to men than to women (Lambrecht and Tucker, 2019), reinforcing gender stereotyping.

However, these loops which reinforce discrimination are best known in the context of the use of AI by police forces, and in particular the development of predictive policing (O'Neil, 2016). Known as a *"feedback loop"*, this phenomenon occurs when police forces are repeatedly sent back to the same neighbourhoods, regardless of the actual crime rate (Ensign et al., 2018). Predictive policing systems extrapolate on the basis of historically and humanly biased data, but these biases are reinforced by the operation of the system itself, which necessarily finds more crime in the areas most covered by the police. The same information is then used to send more police forces to the same area, feeding the loop. While the use of a ML model can be called into question in this case, it is also the use of the model's output as input, to make a new prediction, that causes the system to collapse.

The question of which data to integrate also becomes crucial when the system is designed to learn continuously from new data. This was the case, for instance, of the chatbot Tay, developed by Microsoft in 2016 and deployed on Twitter to interact with web users. However, the system is infamous for sparking controversy after posting offensive messages (Hunt, 2016). The bot is said to have become misogynistic and racist after users forced it to search the internet for sources of its replies, with the hate speech it broadcast being inspired in part by previously tweeted content.

But beyond biases, the opacity of ML models can also have an impact on individuals, specifically when it is reinforced by the general opacity of the ecosystems in which ML models operate. In the Netherlands, the SyRI system, used by the Dutch government to detect social security, benefit and tax frauds, was brought to justice in

[48] See Chapter 5, Section 5.4.1.

2020 (The Hague District Court, 2020). The court of The Hague found that the SyRI system presented potential risks of bias against people of lower socio-economic status or with an immigrant background, a condition reinforced by the fact that SyRI was only deployed in poor neighbourhoods. But what really motivated the Court's decision to declare SyRI in breach of the European Convention on Human Rights was the general lack of transparency in the operation of the system (van Bekkum and Borgesius, 2021). The proportionality and subsidiarity of the SyRI system were assessed and the court found that it was not *"necessary in a democratic society"*.

Safety and security

Another danger of AI is its susceptibility to attacks. Cybersecurity has therefore been a central concern of the AI community for some years now, and took centre stage when researchers discovered that it was possible to deceive image recognition systems by introducing noise into the data[49] (Goodfellow et al., 2015). The same trick could be used to turn stop signs on the road into simple speed limits, which could lead to incidents (Eykholt et al., 2018). With GenAI, the cybersecurity risk continue to be present, as new research has shown that ChatGPT – and probably other LLMs – is vulnerable to many traditional attacks Gupta et al. (2023) and can be exploited to create new attacks Yao et al. (2024), particularly at the user level due to its human-like behaviour.

But even without attack, AI systems are prone to safety issues. The Uber self-driving car crash in 2018 is a case in point, as both the limitations of the computer vision system and human over-reliance on technology were at the heart of the problem (Kohli and Chadha, 2019).

The invisible human and environmental costs

Another challenge of AI is the human cost of training models. To function, AI models have to rely on data annotated by humans. These annotations are produced by data workers, who often work in the Global South, as subcontractors for companies in the North (Ludec et al., 2023). They work in difficult conditions, sometimes without proper offices, on highly repetitive tasks. Wages are extremely low, sometimes less than a dollar per hour. Many data workers are overqualified for this type of job, but they no other prospects. They suffer from being invisible to the general public, which leads to low self-esteem compared with data scientists and engineers, who are often promoted by subcontracting companies (Ludec et al., 2023). In addition, some of these workers are tasked to review and annotate harmful content, with little to no consideration over their mental health (Perrigo, 2023). For Ludec et al. (2023), with data labour, we are witnessing a *"long-standing dynamic of job displacement"*, with the explosion of cheap and tedious labour to produce AI solutions.

Finally, the materiality of AI is not only apparent when it comes to data labour, but also when it comes to its environmental cost. In their study, Strubell et al. (2019) showed that the process of training a large language language model, with a transformer architecture, can lead to approximately 280,000 kg[50] of carbon dioxide emissions, representing what an average human being produces in more than 50 years of life[51]. This finding has given rise to a trend of studies around "sustainable AI", seeking to reduce the environmental footprint of the production and use of AI models. While many studies focus on the development of "green" or "energy-

[49] These attacks are called adversarial attacks.

[50] Translated into kg from the estimate in lbs in the original study: approximately 620,000 lbs.

[51] An average human being on Earth produces 5,000 kg – or 11,000 lbs according to the study – of carbon dioxide emissions in one year. Of course, this is figure is way higher in Europe and in the US, with an American human being producing closer to 16,000 kg a year (Strubell et al., 2019).

efficient" AI (Bolón-Canedo et al., 2024), for van Wynsberghe (2021), sustainable AI should aim to foster change in the entire lifecycle of AI products, not only based on technical improvements, but also considering the social dimension of AI.

These social dangers will not be easily resolved in the near future. However, several partial solutions exist to address these dangers, including sociological critique of AI or philosophical reasoning. In this thesis, we focus on regulation as one of the possible responses to these dangers.

1.1.7 *Conceptualising regulation and its mechanisms*

What is regulation?

With the growing number of challenges associated with AI, it is necessary to address and control them. These attempts can be seen as forms of regulation. To regulate is broadly understood as *"to govern or direct according to rule"* (Merriam-Webster Dictionary, n.d.i). Another way of defining regulation would be to list the instruments and means available to various entities to enforce this regulation. To be able to regulate, an entity must not only have the capacity to govern according to rules, but also the power to create and propagate these rules (Brunsson and Jacobsson, 2002). For Hood et al. (2001), this includes the capacity for standard-setting, information-gathering and behaviour modification. For this thesis we will take the following definition.

> ***Definition***
>
> **Regulation.** *The rule-making and governance capacity of an entity or group, where rules can be enforced through a variety of instruments.*

In this definition, there is no limitation as to who can – or should – exercise this form of governance and rule-making. However, when we think of means of regulation, we think first and foremost of mandatory "hard" law instituted by the *"nation state"* (Morgan and Yeung, 2007). Indeed, we are more familiar with state regulation, because the functioning of our modern society is dictated by the law that we all recognise as citizens. What differentiates legal regulation from other forms of regulation, is that laws are authoritarian rules, with a legitimate – and ideally democratic – source of enforcement power. Morgan and Yeung (2007) recognises the diversity of definitions of regulation, stating that: *"At their narrowest, definitions of regulation tend to centre on deliberate attempts by the state to influence socially valuable behaviour [...]. At its broadest, regulation is seen as encompassing all forms of social control, whether intentional or not, and whether imposed by the state or other social institutions."* Indeed, we often fail to recognise other forms of regulation that are not centred on the state, but are found all across our society, what Black (2002) calls *"decentered regulation"*. Similarly, Abbott and Snidal (2009) speaks of new forms of regulation beyond state regulation: regulation *"above"* the state, in the form of international regulation, and regulation *"below"* the state, with efforts coming from various stakeholders and the civil society. This "below" the state regulation seeks to challenge state regulation and propose alternative means of control to binding law. However, it is often motivated by business incentives. These forms of non-state regulation are increasingly recognised as legitimate and useful means of regulation, complementing traditional regulation (Hutter, 2006). Therefore, "regulation" does not necessarily means "mandatory legislation". It can come from a number of different entities.

The possibility of regulation emanating from society implies a transfer of power from the top down, from states or supra-state entities to groups of stakeholders which are normally those being regulated, such as civil organisations or private companies. However, since this power is not fully – but only partially – transferred, this raises questions about the relationship between the state and these non-state regulatory entities (Morgan and Yeung, 2007).

Forms of regulations in Europe

Over the last thirty years, the European Union has developed various means of regulation, diversifying regulatory instruments with a view to enhancing the effectiveness and legitimacy of the system (Senden, 2005). These new instruments often fall under three categories: soft law, self-regulation and co-regulation. Soft law is the more evasive and difficult of the three concepts to draw. It usually encompasses everything that is not "hard law", i.e. legally binding obligations (Abbott and Snidal, 2000). As such, soft law can contain a myriad of regulatory instruments. Some scholars have argued that soft law represents a danger to the normative system, rendering it inoperative (Weil, 1983). But soft law can also serve as an intermediate step towards binding law, or even complement it, as it is more flexible and therefore easier to develop and implement (Abbott and Snidal, 2000). For Senden (2004), soft law can serve as *"pre-law"* – contributing to the preparation of a legal act, post-law – to influence its application, or *"para-law"* – to provide for complementary rules.

Self-regulation is defined by the European Union as *"the possibility for economic operators, the social partners, non-governmental organizations or associations to adopt amongst themselves and for themselves common guidelines at European level"* (European Parliament, Council and Commission, 2003). Similarly, co-regulation is defined as *"the mechanism whereby a Community legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognized in the field (such as economic operators, the social partners, non-governmental organisations, or associations)"* (European Parliament, Council and Commission, 2003). Self-regulation and co-regulation share a common feature: they bridge the gap between binding legislation and soft law (Senden, 2005). But whereas self-regulation is a *"bottom-up"* approach, where the private sector sets its own rules, co-regulation is seen as a *"top-down"* approach, where public and private entities work together to define these rules, but where public institutions remain in the driving seat (Best, 2003). The state or supra state entity intervenes during the agenda setting, while the implementation is left to private actors (Abbott and Snidal, 2009). On European level, this implies that the European legislator sets the legal framework and the parties involved, and that an institution, such as the Commission, monitors the result and sometimes transforms it into binding legislation (Senden, 2005). Almada and Petit (2023) even calls it *"meta-regulation"* as the European institutions only monitor the self-regulation of private companies. Co-regulation is therefore an intermediary between pure self-regulatory mechanisms set up by and for private entities and pure binding legislation set up entirely by governmental entities. For Abbott and Snidal (2001), this blend of public and private governance is necessary to regulate properly, by mobilising the specific competencies of these two sectors. For the European

Commission (2001), co-regulation makes it possible, in particular, to draw on the practical expertise of the actors most concerned. However, for Streeck (1995), this policy strategy is proof that the EU's capacity to impose binding regulations itself is weak, and that it is forced to supplement them with soft laws, what they call *"neo-voluntarism"*. A simplistic definition of co-regulation would be based on a duality between public and private entities; however, for Abbott and Snidal (2009), it would be more realistic to view governance and regulation as a triangle, with state, firms, but also civil society and non-governmental organisations. Standards are an example of cooperation between these three types of actors[52].

What instruments to regulate AI?

AI regulation can take the form of state or regional regulation, as is the case in the European Union with the AI Act (European Parliament and Council, 2024b), but also in various other regions of the world[53]. It can also take the form of supra-state regulation, for instance international agreements such as the Council of Europe[54]'s Convention on AI (Council of Europe, 2024a).

But if regulation is understood in its larger sense, encompassing self- and co-regulation, then the instruments of regulation, what Freiberg (2010) calls the *"regulatory toolkit"*, is also broader. In the context of AI, ethics charters, for instance, can be considered as soft law and, when they come from the private sector, a form of self-regulation (de Almeida et al., 2021).

Standards as also widely viewed in the literature as a means of regulation[55]. For Kamara (2021), standards are soft laws which can serve pre-, post- or para-law objectives. At European level, they form part of a larger framework on product safety[56], where private actors draw up standards which are then validated – or not – by the European institutions. European standards are therefore a form of co-regulatory instrument (Kamara, 2017) that emerge from a top-down supervised process (Senden, 2017). Standards can therefore also be used to regulate AI, and efforts in this direction are already underway at national, European and international level. In Europe in particular, they will be used to support the AI Act[57]. AI regulation can therefore be very diverse and ethics charters, the AI Act and technical standards act as three different forms of AI regulation.

1.1.8 Difficulties in regulating AI

Different conceptions of regulation

Despite the dangers it poses to society and the apparent need to control it, efforts to regulate AI face a number of obstacles linked to the technology itself and its ecosystem. First, there are different conceptions of what AI regulation would look like. Benbouzid et al. (2022) identify four regulatory arenas in which different actors have different visions of AI and how it should be controlled.

The first two arenas focus on regulating AI as a scientific discipline, while the third and fourth arenas seek to regulate AI as a product or system. The first arena is made up of actors who believe in a dystopian future where we coexist with an all-powerful AI that represents a threat to the human species. Regulation therefore focuses on *"existential risks"* and on how to anticipate the moment when the power of the machine will escape human control. The second arena focuses on the

[52] For more information, see Chapter 4, Section 4.4.1.

[53] We discuss other initiatives and how the AI Act has influenced them in Chapter 9, Section 9.7.3. For an overview of state and supra-state regulation initiatives worldwide, see (White & Case, n.d.).

[54] The Council of Europe is an international NGO with members outside of Europe.

[55] For more information, see our background study on technical standards as a means of regulation in Chapter 4, Section 4.4.3.

[56] For more information on product safety in Europe and the role of technical standards, see Chapter 4, Section 4.6.

[57] For more information on the interaction between the AI Act and technical standards, see Chapter 4.

responsibility of researchers to find solutions for controlling AI. For this sphere of actors, research into technical shortcomings can help control AI and prevent harm by aligning technology with a set of principles, such as fairness, explainability, privacy, which are considered technical properties of AI systems. Actors in the third arena look at AI from the angle of social critique, focusing on regulatory means that protect the fundamental rights of individuals and our democratic values. Finally, actors in the fourth arena see AI as a technico-economic system, where regulation is achieved through product compliance. This is notably the sphere of the European institutions.

The stakeholders in these four arenas occasionally exchange ideas, but more often function as different social spheres, each proposing actions that add to each other rather than work together. As a result, AI regulation takes many forms, and it is difficult to coordinate efforts. In the European context, the AI Act falls mainly within the arena of compliance, although it claims fundamental rights aspects that are closer to the arena of socio-technical critique. In addition, it proposes a number of requirements for high-risk AI systems akin to technical alignment. This mix of approaches makes it difficult for the European text to align means and ends (Smuha and Yeung, 2024), and sometimes gives the impression of a patchwork of regulatory measures.

Different technologies and contexts

Another difficulty in regulating AI is that the term itself is highly polysemous: it encompasses a variety of different technologies used in different contexts. The operation and structure of two AI algorithms can be fundamentally different – for instance between symbolic AI versus machine learning, supervised versus unsupervised learning, traditional ML versus deep learning, and so on. These various systems each have their own technical shortcomings. But they can also be used in a variety of contexts and applications, each presenting unique dangers to individuals and society. Therefore, different AI technologies used in different contexts require different degrees and means of control.

Although the AI Act (European Commission, 2021e) is a horizontal regulation designed to cover all types of AI technologies, it takes account of this diversity in the different categories of systems it covers. In particular, AI systems will be classified as "high-risk" according to their use case – health, employment, border control, and so on. However, the requirements applicable to these high-risk systems will be harmonised across domains[58].

This horizontality poses two main problems. Firstly, it is difficult to define requirements that are broad enough to cover all situations, while ensuring that they can be implemented in practice. While the AI Act stays at a high level when defining these requirements, in order to apply them, producers of AI systems will have to interpret them in their own context to decide which technical solution to choose. Secondly, this horizontal regulation will have to co-exist with sector-specific vertical regulations. However, multiplying the levels of regulation runs the risk of having incoherent obligations between horizontal and vertical regulations.

[58] Reasons for this approach are given in Chapter 3, Section 3.2.4. See also Chapter 3, Section 3.5 for more information about the content of the AI Act.

Complexity of AI systems

AIs are highly complex systems which, apart from the difficulty of defining them and the myriad of technologies they encompass, have a number of intrinsic properties and technical shortcomings[59], making it difficult to impose strict rules on their operation[60].

For instance, AI systems are based on correlation rather than causation. This has led some authors to claim that the creators of AI themselves do not know how it works. This is not entirely true insofar as the developers of AI systems know the technology and its limits, but cannot anticipate the result of an AI computation because of the large number of parameters and possible outcome. Due to their probabilistic nature, these systems learn new rules, but it is currently impossible to access which rule a model has learned and why. It is therefore virtually impossible to certify that, in a given situation, a certain result will be produced. This lack of verifiable metrics is particularly detrimental when AI is deployed in a variety of situations. For example, an AI-based radar could have different levels of performance in sunny or rainy weather, making real-world testing almost pointless (Ebers, 2022). This is reinforced by the opacity of AI models, which makes it difficult to understand their behaviour and ensure that the model comply with regulatory requirements.

Moreover, another difficulty linked to the regulation of AI technologies lies in the possibility that they will continue to learn once deployed in the context of continuous adaptation (Ebers, 2022). While not all models have this feature, for those that do, it is a daunting task to keep track of all versions of the model, their rules and evolving behaviour, and to ensure that regulatory requirements are met at all times.

A rapid evolution of the technology

AI is often considered to be a recent technology. However, its history shows that it has its roots in the 1950s, and that we have even known how to train neural networks since the late 1980s. But the explosion in AI performance is much more recent, with deep learning only overtaking classical symbolic models in the 2010s. Another example would be GenAI, which, although the technology existed before ChatGPT and Midjourney, has only been used by the general public since the end of 2022, about two years ago. The technology therefore evolved very quickly and it is complicated for regulation to keep up the pace. As technology evolves, so do the ethical issues it raises, requiring a rethink of existing frameworks. The AI Act, for instance, incorporated a new category of systems, General Purpose AI (GPAI), along the way, to account for the rise of GenAI. The more technical the framework, the more difficult it is to lay down rules and requirements, as the technical details are likely to become obsolete very quickly.

Moreover, creating effective technical requirements requires practical experience, which even the AI deployers have very little of (Pouget and Laux, 2023). Indeed, we are only at the beginning of the deployment of this technology and it is difficult to anticipate all the impacts it will have on individuals and society.

[59] See Section 1.1.5.

[60] Ebers (2022) explains these difficulties in the context of AI standards.

Translating requirements

Finally, AI is a multi-hands technology (Edwards, 2022). Indeed, AI systems are complex socio-technical systems in which a technological object is defined, built, deployed, regulated or used by a set of different actors, with their own background, experience, vocabulary and personal opinions, who interact as individuals or through institutions[61]. Throughout its lifecycle, AI will pass through different hands: those of producers, contractors, legal consultants, deployers, users and operators, regulators, data subjects, or subjects of an algorithmic decision... It is therefore difficult to know which obligations fall to which person, especially when the production phases are not linear nor independent. But it also requires a great deal of effort in terms of comprehension between stakeholders, as instructions are passed from hand to hand, with a lot of information getting lost in translation.

Part of this thesis therefore focuses on how regulation moves from ethics to mandatory law and technical requirements, and how it changes what we impose in terms of AI control.

1.2 Thesis overview

This section is used to shed light on how the thesis is constructed, from the problems it addresses to how it addresses them, chapter by chapter.

1.2.1 Problem statement

It is not an easy and straightforward task to define the terms associated with AI, the properties these systems must respect or the means of controlling them. Yet, with the rise of AI technology and, with it, the emergence of numerous dangers to individuals and society, many stakeholders have sought to develop initiatives to regulate AI. From the first AI ethical charters to the legal framework of the AI Act, it has become clearer that everyone wants AI to be "trustworthy". But what "trustworthy" means has yet to be agreed upon[62].

In Europe, specifically, the emergence of the AI Act created a need to define more precise legal and technical requirements. The law follows a discourse on the protection of fundamental rights through a risk-based analysis strongly inspired by ethical documents. But at the same time, the AI Act is based on an approach to product safety that favours the use of technical standards to address these legal and ethical issues, a tool that is not designed for this purpose. There seems to be a flow of horizontal, high-level ideas that have appeared in AI ethics charters, been retained in the AI Act and are now finding their way into the technical standards that are supposed to support this legal text. Therefore, the question of how different regulatory frameworks fit together is becoming all the more important.

At the same time, an in-depth examination of these standards and the functioning of their ecosystem reveals a large number of shortcomings which, although already present in the traditional product safety approach, are reaching new heights with AI. As the economic model of standards is increasingly being called into question, it is in fact the fitness of the current system as a whole to regulate AI that could be challenged.

In this thesis, we examine the European approach to AI regulation and we aim at answering one main research question:

[61] This idea that different actors gravitate around a socio-technical object and fight to impose their vision and terminology is not new in Science and Technology Studies (STS), and especially in sociology of translation. See for instance, the work of Callon (1984)

[62] The term "trustworthy AI" itself is the subject of debate. It is sometimes replaced by "responsible" or "ethical AI" in a semantic battle between different institutions. The European Union has chosen the term of "trustworthy AI" since the work of the High Level Expert Group on AI (HLEG), and has continued to use it in its communications and documents, and the term has notably being reused in standards. For more information on the "truswor-thiness" appraoch of the EU, see Chapter 3.

Research Question – What makes Europe's approach to regulating AI so distinctive from other European regulations?

To answer this Research Question we will focus on two objectives:

Research Objective 1 – Understand the interplay between AI ethics charters, regulation and standards. We will examine each of these three types of documents, with their respective ecosystem, deciphering their relationship and examining the issues that might arise when conflating ethics, law, and technical requirements.

Research Objective 2 – Comprehend the world of standardisation, the content of standards, their development process and the obstacles they face. We will analyse current initiatives, looking at who is working on which topics and the potential dangers of the current approach to regulating ethics and fundamental rights through technical tools designed mainly by the industry.

1.2.2 *Structure*

This thesis is divided into ten main chapters, and two main parts. Chapter 1 is an introduction to the thesis and Chapters 9 and 10 contain the discussion and conclusion. Between them, the core of the thesis is concentrated in Chapters 2 to 8, which present the main findings of the research. Each chapter can be read independently, as the concepts needed to understand the main ideas of the chapters are generally recalled in the first sections.

INTRODUCTION

Chapter 1 – Introduction to the regulation of AI: definitions and thesis scope
The first chapter of this work serves as an introduction to AI and regulation. It tells the story of AI, situates AI in the fields of computer science and in law, highlights the challenges associated with its use, both in terms of technical shortcomings and the dangers for individuals and society as a whole. This introduction also highlights the need for AI regulation, the different types of regulation that exist as well as some examples of AI regulation instruments.

PART I – A THREE-STEP APPROACH TO AI REGULATION
The three chapters in this first part each focus on one of the three instruments which we study in this thesis: AI ethics charters, the AI Act, and technical standards. They can be seen as an introduction to each of these instruments, presenting in particular their development and content. However, in addition to the necessary background elements, they each bring their own original results and should therefore be considered as an integral part of the thesis contributions.

Chapter 2 – AI ethics charters: a tension between regulatory, technical and business discourses
This chapter adopts a broad discursive study angle, examining all initiatives on the ethics of AI. It differs from previous meta-analyses by adopting a quantitative approach. In this work, we map 436 initiatives around AI ethics, creating a corpus of documents that will help foster future research. We further examine the similarities in their discourses, showing that certain themes – such as fairness – have permeated the sphere of AI ethics, being used by a wide range of actors and with a wide range of vocabularies. Interactions between a wide range of stakeholders could explain this diffusion and convergence of ideas, steering the general discourse on the ethics

of AI in the same direction: that of broad principles that are difficult to implement in practice.

Chapter 3 – The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation

This chapter recounts the story of the AI Act, from the first texts on AI ethics in Europe to the mandatory regulation we know today. This chapter also serves as a starting point for understanding the scope of the AI Act, the main categories of systems it covers and the enforcement mechanisms it puts in place. We show that the approach set out in the AI Act to regulate through risk, while relying heavily on a discourse on "trustworthiness", is inherited from a long history of policy documents and ethics charters in Europe. At the same time, the AI Act aims to establish a horizontal framework, drawing from the General Data Protection Regulation (GDPR), while incorporating a vertical product safety approach.

Chapter 4 – Technical standards: a product safety approach adapted to AI

This chapter serves as an introduction to standardisation. It shows how difficult it is to define what a standard is, and outlines the main roles that standards can play, before focusing on the specific European context. In the European Union, standards are used for safety purposes, to assess the conformity of products with harmonised legislation. But these standards come from a specific ecosystem with its own complex rules. In addition, this chapter shows how the AI Act fits into this new legislative framework for product safety in the EU, explaining the role standards and conformity assessment will play for AI.

Chapter 6 – Case study: fairness standards, the new ethics charters

This chapter dives deeper into one use case: the principle of fairness. It analyses how it is defined in main ethics charters and the AI Act, as well as lists current standardisation initiatives which deal with this issue. It shows which aspects of fairness are well covered and where there are gaps to be filled in future initiatives. We conclude that some aspects of fairness have been lost in the transition from ethics charters to the AI Act and subsequently to standards, with fairness being reduced to data quality. We also show that fairness standards remain at a very high-level, not being necessarily more technically specific than ethics charters.

PART II – A FOCUS ON STANDARDS TO REVEAL TENSIONS BETWEEN REGULATION MECHANISMS

Chapter 5 – Standardising ethics and fundamental rights: the difficult task of AI standards

This chapter discusses the new angle adopted by the EU with the AI Act: the potential standardisation of ethics and fundamental rights. It shows how the AI Act introduces this shift, what initiatives are currently underway in this area and the problems it may pose. Finally, it provides guidance on what standards could address and what they should probably avoid.

Chapter 7 – Regulation through standardisation: the legitimacy issue of European AI standards

In this chapter, we dive into the principal issues posed by the European standardisation system. In particular, we examine the status of harmonised standards under EU law in the light of recent case law on access to standards. We show that the current standardisation system was already criticised, even before it was applied to

AI. But with the AI Act, the problems of European sovereignty, regulatory capture and the democratic legitimacy of the world of standardisation have reached a new level, forcing the European Commission to prepare a safety net in case standards fail.

Chapter 8 – Behind the scenes of AI standards-making in Europe: interviews with CEN-CENELEC JTC 21 experts

After having examined the standardisation system from the outside, we turn to the standardisation experts to tell us directly about their work. In addition to a quantitative analysis of interest representation in standardisation, we also interviewed 16 experts from CEN-CENELEC JTC 21 who are working on the future European AI standards, and asked them about the organisation of their work, the difficulties they face and the reasons for their involvements. Their testimonies highlight the limitations of standardisation work, from staff shortages to the industry's stranglehold on standards. The study shows that AI standardisation is a highly political field, far removed from technical considerations. The experts also describe possible improvements to the current system.

DISCUSSION AND CONCLUSION

Chapter 9 – Discussion: regulating AI through vertical sectors

This chapter aims to bring together the many facets of AI ethics, regulation and standardisation that we studied throughout the thesis. It serves as a general discussion, recapitulating what we have learned and linking it to the shortcomings of AI regulation today. It highlights the need to refocus AI regulation on vertical sectors. Indeed, the discourse of AI ethics charters has influenced the AI Act and technical standards, each time remaining at a very high and horizontal level. But for the AI Act implementation, precise technical requirements will have to be defined, and this can only happen if standards are sector-specific.

Chapter 10 – General conclusion

This chapter concludes the thesis, summarising our findings and giving final recommendations. It also suggests a few avenues for future research, and provides advice for researchers wishing to delve deeper into the various facets of AI regulation.

1.2.3 *An interdisciplinary work*

I believe that it is impossible to understand the challenges of AI without interdisciplinarity. Even before I started this PhD, I had encountered AI across a number of disciplines. I am first and foremost an engineer, trained in AI and machine learning from a technical standpoint. I also studied AI during my Master's degree at Sciences Po, through the prism of political science. Finally, my internship at the French Digital Ethics Committee (CNPEN) gave me an insight on the ethics of AI. The work of this thesis draws on all these disciplines and goes even further.

This thesis is presented in "Law and Regulation" and therefore it necessarily draws on legal studies for many of its contributions (notably in Chapter 3, Chapter 4 and Chapter 7). But many other disciplines have influenced me for this work.

First and foremost, although no contribution is made specifically in the field of computer science, influences from this discipline can be seen throughout the thesis. First of all, data science and data analysis have been essential to several of my contributions (Chapter 2, Chapter 6 and Chapter 8). But most importantly, the understanding of AI and machine learning from a technical standpoint was crucial to my research and my interpretation of the results. In particular, it has enabled me to develop my arguments using technical examples such as fairness, which is often mentioned and analysed in this thesis (Chapter 5 and Chapter 6).

Finally, this thesis would not be complete without its contributions inspired by the social sciences. These include computational social sciences (Chapter 2) and qualitative sociology (Chapter 8).

I believe, however, that this manuscript should be read bearing in mind that the author's views are those of an engineer analysing legal texts, regulation mechanisms, and social spaces, rather than those of someone with specific training in legal studies or social sciences. I strongly believe that this prism gives my research a unique perspective on AI regulation, one that I personally hope future researchers will adopt, because the field of AI regulation, while in great need of legal experts and social scientists, is also in need of technical experts.

1.3 Publications and other productions

This section lists all the publications and concrete productions resulting from this thesis. It is separated into academic publications and productions – work published as a first author, co-authored publications, and preprints – and other types of documents – chapters in books, online articles and communications. In each sub-section, the publications are presented in anti-chronological order.

Finally, I present some of the related activities I conducted during my PhD – scientific posters, invited talks, teaching support, etc.

An asterisk (*) indicates that the article is used in this manuscript. The parts in which this work appears are indicated in brackets. At the beginning of each chapter, a short paragraph explains when an article has been used, what changes have been made and why. None of these publications have been used *as is*. Chapter 2 takes one article and extend it by proposing a more complete background analysis, new results and a new discussion. One other article has been split and parts of it are used in several chapters. For some other chapters, some sections of these articles are re-used, with the appropriate reference at the beginning. However, much of the work and writing in this thesis is new. This thesis should therefore be seen as a new study, with its own questions and its own demonstration.

1.3.1 Publications as a first author

Academic journals

*Gornet, M. and Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, 13(3). `https://doi.org/10.14763/2024.3.1784` (used for Chapters 4, 5 and 7)

International conferences

*Gornet, M., Delarue, S., Boritchev, M. and Viard, T. (2024). Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 127-140, Rio de Janeiro, Brazil. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658545 (used for Chapter 2)

National conferences

Gornet, M. and Maxwell, W. (2023a). L'IA explicable appliquée à la détection de ceintures et de téléphones au volant. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA @ PfIA 2023)*, pages 46-56, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). https://hal.science/hal-04158889/

*Gornet, M. and Maxwell, W. (2023b). Normes techniques et éthique de l'IA. In *Conférence Nationale en Intelligence Artificielle (CNIA @ PfIA 2023)*, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). https://pfia23.icube.unistra.fr/conferences/cnia/publications/Gornet.pdf (used for Chapter 5)

1.3.2 Co-authored publications

Herman, H. and Gornet, M. (2024). La normalisation de l'IA: un déluge de réinterprétations de l'AI Act. In *Journée Société et IA (SIA @ PFIA 2024)*. Association Française pour l'Intelligence Artificielle (AFIA). https://pfia2024.univ-lr.fr/Pr%C3%A9sentations/Soci%C3%A9t%C3%A9-et-IA---Pr%C3%A9sentation-3.1/

Viard, T., Gornet, M. and Delarue, S. (2023a). Interpretive flexibility in data science and artificial intelligence. In *Undone Computer Science*. Nantes, France. https://undonecs.sciencesconf.org/data/Undonecs_2024_abstract_11.pdf

Viard, T., Gornet, M. and Maxwell, W. (2023b). Reading the drafts of the AI Act with a technical lens. In *Regulatable ML Workshop @ Neurips 2023*. https://openreview.net/pdf?id=pnvRy1VzJZ

1.3.3 Preprints

Preprints resulting from this thesis

Gornet, M. (2024a). The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation. https://hal.science/hal-04785519 (resulting from Chapter 3)

Gornet, M. (2024c). Regulation through standardisation: the legitimacy issue of European AI standards. https://hal.science/hal-04785178 (resulting from Chapter 7)

Gornet, M. and Herman, H. (2024). A peek into European standards making for AI: between geopolitical and economic interests. https://hal.science/hal-04784035 (resulting from Chapter 8)

Gornet, M. (2024d). Too broad to handle: can we "fix" harmonised standards on artificial intelligence by focusing on vertical sectors? https://hal.science/hal-04785208 (resulting from Chapter 9)

Other preprints

Gornet, M. and Viard, T. (2023b). Queer identities and Machine Learning. https://hal.science/hal-04763352

Gornet, M. and Viard, T. (2023a). Mapping AI Ethics: A Quantitative Analysis of the Plurality, and Lack Thereof, of Discourses. https://dx.doi.org/10.2139/ssrn.4582657

Gornet, M., Kirchner, C. and Tessier, C. (2022b). Operational fairness when coding facial authentication. https://hal.science/hal-04447868v1/

1.3.4 Other productions

Chapters in books

Gornet, M. Limitations techniques du machine learning et choix de conception. In Nevejans, N. *Décision humaine, décision de l'IA*, to be published.

Tessier, C., Chatila, R., Devillers, L., Dognin-Sauze, K., Ganascia, J.G., Gornet, M., Pronesti, A. (2022). Pourquoi la reconnaissance faciale, posturale et comportementale soulève-t-elle des questionnements éthiques ? In Germain, E., Kirchner, C., Tessier, C. *Pour une éthique du numérique*, PUF.

Communications

Gornet, M. and Maxwell, W. (2023c). Recherche sur les normes techniques pour l'IA et les droits fondamentaux à Télécom Paris. Bulletin N°120 "IA & Normes". pages 14-19. Association française pour l'Intelligence Artificielle (AfIA). https://www.hub-franceia.fr/wp-content/uploads/2024/03/AFIA-bulletin-120_avr23.pdf

Gornet, M., Kirchner, C. and Tessier, C. (2022a). Operational Fairness for Facial Authentication Systems. *ERCIM News N°131*, Special theme on Ethical Software Engineering and Ethically Aligned Design. `https://ercim-news.ercim.eu/en131/special/operational-fairness-for-facial-authentication-systems`

Online articles

Gornet, M. (2024b). La stratégie européenne de régulation de l'intelligence artificielle : entre normes techniques et droits fondamentaux. Série de blog "Voix émergentes dans l'IA et la société", University of Ottawa. `https://www.uottawa.ca/recherche-innovation/toutes-nouvelles/strategie-europeenne-regulation-lia-entre-normes-techniques-droits-fondamentaux`

Gornet, M. and Maxwell, W. (2022). Intelligence artificielle : normes techniques et droits fondamentaux, un mélange risqué. *The Conversation*. `https://theconversation.com/intelligence-artificielle-normes-techniques-et-droits-fondamentaux-un-melange-risque-189587`

1.3.5 *Other activities related to my PhD work*

Scientific posters

I produced and presented two scientific posters during my PhD.

The first one, *Standardizing fairness for AI systems*, was presented at the 2022 doctoral school event and won the **IP Paris PhD Poster Award** in the field of Economics, Management, and Social Sciences.

The second one, *Mapping AI Ethics: a quantitative analysis of the plurality, and lack thereof, of discourses*, was presented in Télécom Paris during the 2023 "company day".

Invited talks

In addition to conferences and workshops where I have presented my academic publications (FAccT, CNIA, uOttawa, etc.), I have also been an invited speaker to a number of events. Here is a selection of my talks:

- *What regulation for AI in Europe? The AI Act, technical standards and fundamental rights*. Semaine de l'IA, Université de Rennes, 3 Septembre 2024.
- *Technical introduction to AI and machine learning.* Legal Workshop: The Digital Transformation of the Public Sector, University of Basel, 3-5 July 2024.
- *Les limitations techniques du machine learning.* Colloque Décision humaine, Décision de l'IA, Université d'Artois, 8-10 Novembre 2023.
- *Reading the AI Act with a technical lens* (with Tiphaine Viard). Training course on the Artificial Intelligence Act at the Council of the European Union, 10 July 2023.

This list could also be completed by the various roundtables and webinars in which I have participated (chaire IRSN at Ecole Polytechnique, AI Mondays at Télécom Paris, etc.).

Teaching support

I also had some teaching experience during my PhD. Here is a selection:

- Ecole Polytechnique: "Ethical Issues, Law and Novel Applications of AI", MScT AI-VIC, seminar organisation and animation (1h30/ week, 2022-2024)
- ISAE-SUPAERO: "Introduction to data protection", Master 2, lecturer (9h, 2022-2024)
- Institut Polytechnique de Paris: "Law and ethics of artificial intelligence", Master 2, design and animation of a group exercise (10h, 2022)
- Télécom Paris: "Mining and exploring large datsets", Master 2, lab sessions (2022)

I have also been invited to make short presentations during other courses (at Télécom Paris, ISEP, etc.).

Organisation of events

I have helped organised two webinars for the LIMPID project – *Facial recognition: towards a transatlantic normative scheme on trustworthiness?* and *Biases and computer vision*; in collaboration with Paris 1 Panthéon-Sorbonne.

I was also part of the organising committee of the Transatlantic AI and Law Institute (TALI), an initiative by Chinmayi Sharma, in collaboration with Fordham Law School, Paris 1 Panthéon-Sorbonne, the University of Ottawa, Sciences Po Paris and Télécom Paris. TALI is a series of conference-like events, with speakers invited to present their academic papers and workshops. Three TALI events have already taken place since 2024, in New York and Paris, and others are planned for the future.

Science popularisation

In 2024, I participated to the first edition of IP Paris "3 Minute Thesis", also called in French *Ma thèse en 180 secondes (MT180)*, where PhD candidates have 180 seconds to summarise their thesis' work. I have reached the **IP Paris regional final**[63].

I did the same exercise, this time in English, in the "Pitch for Change" competition organised by Technology for Change, and I won **third prize**.

I have also been invited to talk about AI regulation in videos and podcasts from Télécom Paris and IP Paris[64].

Industrial collaboration

In addition to the collaboration with IDEMIA throughout the LIMPID project, I have also collaborated closely with the Laboratoire National de Métrologie et d'Essais (LNE)[65] as part of a hosting agreement.

[63] A recording of my presentation is available online (in French): https://www.youtube.com/watch?v=QOhqGLPe-WQ. A summary of my pitch is also available on Télécom Paris' website (also in French): https://www.ip-paris.fr/actualites/les-10-finalistes-2024-du-concours-ma-these-en-180-secondes.

[64] See for instance *IA : bâtir une éthique de transparence et de confiance*, https://www.telecom-paris.fr/fr/ideas/ia-ethique-transparence-confiance.

[65] The LNE is a reference laboratory for industry in the field of metrology. To know more, visit their website: https://www.lne.fr/en/about-us/missions.

PART I

A THREE-STEP APPROACH TO AI REGULATION

Chapter 2
AI ethics charters: a tension between regulatory, technical and business discourses

Where does this chapter come from?

This chapter extends on the following article:

> Gornet, M., Delarue, S., Boritchev, M. and Viard, T. (2024). Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 127-140, Rio de Janeiro, Brazil. Association for Computing Machinery. `https://doi.org/10.1145/3630106.3658545`

This article is itself a follow-up to a work that is not presented in this chapter:

> Gornet, M. and Viard, T. (2023a). Mapping AI Ethics: A Quantitative Analysis of the Plurality, and Lack Thereof, of Discourses. `https://dx.doi.org/10.2139/ssrn.4582657`

In this first article, we looked at 74 documents broadly related to AI ethics, and constructed an analysis based on Latent Dirichlet Allocation (LDA), an unsupervised topic modeling method where topics are represented by word probabilities. We showed that broad topics of "Societal Impact", "Policy Recommendations" and "Company Governance" are discussed by various actors across different sectors. This first experiment provided us with a basis for reflection on the variety of discourses around AI ethics and their apparent convergence. Based on these initial results, we built the larger study, on 436 documents, which is presented in this chapter.

Section 2.2, on the background of ethics charters and related works, is longer and more detailed than in the original article. Section 2.3 to 2.5 and 2.7 are taken from the 2024 published article and present our main findings. Section 2.6 and Section 2.8 are original contributions to the thesis. Section 2.6 provides a new analysis of AI ethics documents from EU institutions. Section 2.8 links all these results to our previous findings and proposes an interpretation in terms of meso-scale interactions between different types of actors. The writing style of the two original articles was changed from American to British English, to remain consistent with the rest of the thesis. The list of all the documents used in this study, the corpus, and the datasheet are available at: `http://mapaie.telecom-paris.fr`.

Chapter's abstract

The recent years have seen a surge of initiatives with the goal of defining what "ethical" artificial intelligence would or should entail, resulting in the publication of various charters and manifestos discussing AI ethics; these documents originate from academia, AI industry companies, non-profits, regulatory institutions, and the civil society. The contents of such documents vary wildly, from short, vague position statements to verbatims of democratic debates or impact assessment studies. As such, they are a marker of the social world of artificial intelligence, outlining the tenets of different actors, the consensus and dissensus on important goals, and so on.

Multiple meta-analyses have focused on qualitatively identifying recurring themes in these documents, highlighting the high polysemy of themes such as *transparency* or *trust*, among others. The broad term of "AI ethics" and its guiding principles hide multiple disparities, shaped by our collective imaginations, economic and regulatory incentives, and the pre-existing social and structural power asymmetries; through quantitative analyses, we validate and infirm previous qualitative results.

In this paper, we create and present a corpus of charters and manifestos discussing AI ethics through the process of collection and its quantitative analysis using text analysis to shed light on common and distinct vocabularies. Through frequency analysis, hierarchical topic clustering and semantic graph modelling, we show that the charters and manifestos discuss AI ethics along three broad axes: technical documents, regulatory ones, and innovation and business ones. We use our quantitative analysis to back up and nuance previous qualitative results, showing how some themes remain specific while others have fully permeated the space of AI ethics. We document and release our corpus, comprising of 436 documents, charters and manifestos discussing AI ethics. We release the corpus, its datasheet and our analysis[1], to open the way to further studies and discussions around vocabulary, principles and their evolution, as well as interactions among actors of AI ethics, in order to foster further studies on the topic.

2.1 Introduction

The proliferation of documents around the ethics of Artificial Intelligence has been such that several hundred documents have emerged since the early 2010s. These initiatives to guide AI ethics have been lauded around the world for contributing to opening up the dialogue between different stakeholders on AI benefits and risks, and providing tools to measure the ethical outcome of a decisions. They are seen as a stepping stone to developing AI regulation and binding norms (Langlois and Régis, 2021). However, they are also widely criticised for a variety of reasons: their opacity (Bélisle-Pipon et al., 2022), their Western-centrism and claim to universality (Goffi and Momcilovic, 2022), and their polysemy, that oversimplifies complex ethical debates (Jobin et al., 2019; Ryan and Stahl, 2020). Together, they contribute to outlining AI's social world (Becker, 1976), and understanding it helps shed light on the way knowledge is constructed in AI.

This trend has attracted a lot of attention and has led to numerous meta-analyses (Jobin et al., 2019; Fjeld et al., 2020; Tidjon and Khomh, 2022; Zeng et al., 2018; Hagendorff, 2020; Khan et al., 2022; Attard-Frost et al., 2022), in order to identify common themes and tenets. Both individuals and institutions took hold of this growing

[1] https://mapaie.telecom-paris.fr/.

space, making it inherently sub-political (Beck, 1992), i.e. a space where regulations and societal orientations are decided largely outside of democratic spaces. Describing and understanding these spaces, where the actors and institutions are intertwined with competing interests and multilateral interdependencies, is of crucial importance to understand the social processes and disciplines that span them. This knowledge is key in order for citizens to evaluate the legitimacy of the acting structures and their propositions.

Our main goal in this chapter is twofold: using a quantitative lens, we assess and map out the currents that shape the discussions and tension points around AI ethics[2]; we also provide a structured corpus to foster further analyses, and to unify previous works under a common methodology. The core contribution of this paper is the release of our corpus, containing 436 documents, their contents and some metadata. We provide a meso-scale analysis of the social world (Becker, 1976) of artificial intelligence, while comparing ourselves to previous meta-analyses on the topic. This is, to the best of our knowledge, the first publicly available corpus of this kind, and the second-largest existing database on the topic.

The remainder of this paper is as follows: we start by giving a background on AI ethics charters and discussing related works in Section 2.2, and follow directly by describing our corpus's structure and contents in Section 2.3. We then proceed onto a quantitative analysis of this corpus, exploring term frequencies and topic modeling in Section 2.4, and explore the areas of consensus and controversy with semantic graphs in Section 2.5. We finally expose the limitations of our work in Section 2.7 and conclude in Section 2.9.

2.2 *Background*

2.2.1 *The rise of AI ethics charters worldwide*

The first ethics charters regarding digital technologies arose in the years 2010 to 2015, but at the time, the main subject is not "AI" but "robotics". Main charters include for instance the UK Engineering and Physical Research Council (EPSRC) and their Principles of Robotics (Boden et al., 2017), first drafted in 2011, or the French Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene (CERNA) and their opinion on Robotics Research Ethics (CERNA, 2014). After 2015, data-related issues started to emerge, with the forthcoming arrival of the General Data Protection Regulation (GDPR) European Parliament and Council (2016b) in Europe, and the first private companies like Bitkom or Accenture started releasing Guidelines for Big Data Use (Bitkom, 2016) or Principles of Data Ethics (Accenture, 2016). It is also in these years that the word "AI" first appeared in these documents, with the European Economic and Social Committee (EESC)'s Opinion on AI (EESC, 2017), the US National Science and Technology Council various policy documents and national strategy for AI (NSTC, 2016), as well as the first Tenets on AI from Big Tech companies (Partnership on AI, 2016). The number of AI ethics documents really started exploding in 2017, but it is only in 2018, that the wave reached Europe with the first documents being released by the European Council.

By 2018, both public entities and the private sector were racing to draw up guidelines, with the aim of being the first to impose their vision of AI. The industry was quite efficient, with each company drafting its own AI ethics charter. But the

[2] We recognise that the term "AI ethics" is loaded, notably because it shifts discussions towards making AI ethical, rather than its actors and institutions; furthermore, it assumes that AI *can* be made ethical, by ruling out the alternative of not using or sustaining AI. We use it in this article *because* it is the most common term, rather than out of endorsement.

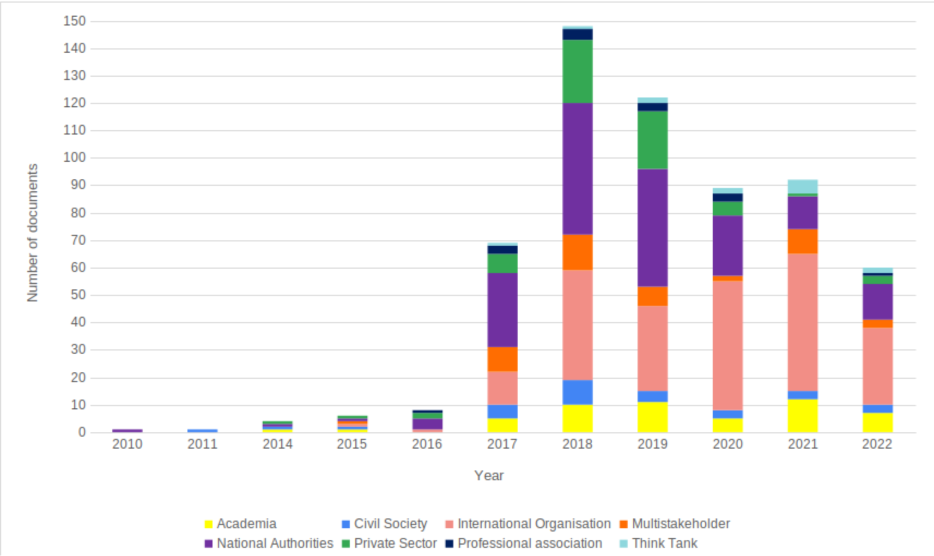


Figure 2.1: Histogram of the number of documents on AI ethics produced per year, by sector. Representation from the author, using data from (Council of Europe, 2022).

number of documents grew even faster for international organisations and national authorities, as they could cover different sectors and each publish numerous documents, whereas companies usually only published one. Figure 2.1 shows this explosion in the number of ethics charters in 2017-2018, with a slight decline in recent years. It shows the diversity of initiatives, which mainly come from the public sector, with national authorities and international organisations in the lead, but also a large proportion of charters emanating from private actors and civil society. Although the ethical principles are fairly similar from one document to another, the differences between the subjects addressed by the various stakeholders around the world were sufficient to cause tensions to arise during the development of the international frameworks, with cultural clashes between countries.

Some prominent examples[3] include the High Level Expert Group (HLEG)'s Ethics guidelines for trustworthy AI (HLEG, 2019a)[4], the Future of Life Institute's Asilomar AI Principles (Future of Life Institute, 2017), the IEEE's Ethically Aligned Design (Chatila and Havens, 2019), The Montréal Declaration for responsible AI development (Université de Montréal, 2018) or the UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021).

2.2.2 *Previous studies on ethics charters*

With the growing trend towards AI ethics charters, it did not take long for the academic world to take an interest. These works include the creation of databases and meta-analyses to examine the content of the guidelines.

Databases

Documents related to AI ethics have become so numerous that there are works dedicated to compiling them. To the best of our knowledge, the most comprehensive database to date is the Council of Europe[5] repository on AI initiatives (Council of Europe, 2022). The database lists over 600 initiatives worldwide, and the collection of new documents stopped in 2022. It is worth noting, however, that the repository contains all sorts of AI related documents beyond simple AI ethics charters, like regulations and laws, technical standards, tools, algorithmic assessments,

[3] We collected, for each document in our database, the number of meta-analyses in which it appears. In particular, this enabled us to see the number of articles covered by each previous meta-analysis. These results are presented in Table 2.1.

[4] The HLEG is a group of experts set up by the European Commission specifically to draw up AI ethics guidelines and related documents. See Chapter 3, Section 3.2 for more information on the content of the HLEG's guidelines.

[5] The Council of Europe is an international NGO, not to be confused with the European Council, which is a European institution.

checklists and other pieces of documents. Another listing, which focused more on ethics charters is the Algorithm Watch inventory of AI Ethics Guidelines (Algorithm Watch, n.d.). Over 160 documents are listed, and in various languages, however, the collection also stopped in 2020. There are also individual initiatives by people who have decided to compile such texts themselves. These include Alan Winfield's blogpost (Winfield, 2019a), presenting a selection of texts with their corresponding principles; the "EthicalML GitHub" (EthicalML, n.d.), pointing to various initiatives, from ethical charters and policy documents, to technical tools and standards; and the AI Ethics Lab's "Toolbox: Dynamics of AI Principles" (AI Ethics Lab, 2020), which features an interactive display where documents and their principles can be explored from a world map.

Others have specialised in compiling a certain type of documents. For instance, the Organisation for Economic Co-operation and Development (OECD) AI Policy Observatory presents a repository of policy papers and national strategies (OECD, n.d.). The fast.ai initiative, founded by two computer scientists and researchers, also feature a web page on "AI Ethics Resources" which points to academics and institutes to follow (Thomas, 2018). These various types of AI-related documents are more and more analysed in specialised meta-analyses, such as Morley et al. (2020) who analyses a whole variety of documents or Durmus (2021) who studies various national AI strategies.

It should be noted, however, that the term "charter" is itself difficult to define and encompasses several different realities. This is one of the difficulties in studying this coalescent object which dominated AI regulation before the 2020s.

Meta-analyses

Additionally to databases, several studies have analysed AI ethics charters in search of common principles for AI, they are called "meta-analyses". The most well-known of these meta-analyses is from Jobin et al. (2019), and studies more than 80 documents published through 2019. They found, through qualitative reading, that five principles were present in more than half of the documents: *transparency*, *justice & fairness*, *non-maleficence*, *responsibility* and *privacy*.

Since then, several other works have explored a similar corpus of texts to identify common topics related to AI ethics. These common topics and principles are either manually found and counted in the texts (Jobin et al., 2019; Fjeld et al., 2020; Tidjon and Khomh, 2022; Zeng et al., 2018; Hagendorff, 2020; Khan et al., 2022), or defined upstream before being searched for in the documents (Floridi and Cowls, 2019; Attard-Frost et al., 2022). Some names may differ, but scholars seem to agree at least partially on the major themes present in the texts. Recurring themes that are present in all the meta-analysis are, in no particular order: *privacy*, *transparency*, *fairness*[6], *accountability*[7], and *safety*[8]. Other themes are less common and cited only in some studies, like *well-being*, *human oversight*, *solidarity*, *explainability*, *collaboration*... However, studies do not always agree on which principle is the most recurrent. *Transparency* is the number one principle in Jobin et al. (2019); Tidjon and Khomh (2022), but for Fjeld et al. (2020); Hagendorff (2020) it is *privacy*.

However, the definition of these principles is hard and alignment cannot be guaranteed from one meta-analysis to another. For instance, *transparency* is very frequent in AI ethics charters according to the work of Jobin et al. (2019), but this is partly because every document containing generic terms such as *"disclosure"* or *"showing"*

[6] The principle of *fairness* is also referred to as *justice* or *non-discrimination*.

[7] The principle of *accountability* is also referred to as *responsibility*, even though the two notions have different meanings. We will consider here that they belong to the same broad theme, since we are trying to group together rather than separate.

[8] The principle of *safety* is contained in the principle of *non-maleficence* in Jobin et al. (2019), and sometimes also grouped with *security*.

was considered to discuss the principle of *transparency*. Other studies use different keywords and therefore have also different results.

Instead of identifying these principles in the texts, some studies begin by establishing principles that they consider the best set of what constitutes "ethical AI". Notably, Floridi and Cowls (2019) builds a set of common principles around the four core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice*, to which they add a new principle, specifically for AI ethics: *explicability*.

Additionally, Fjeld et al. (2020) offers an overview of the distribution of these themes among the documents according to their sector: civil society, government, private sector, government, intergovernmental organisation, and multi-stakeholders. Furthermore, Zeng et al. (2018) gives an analysis of the frequency of topics mentioned across sectors. For instance, the principles of *privacy* and *security* are mostly cited by governments while *humanity*[9] and *accountability* are mostly cited by academia.

Tidjon and Khomh (2022) conducts a similar analysis according to the documents' countries of origin. They note that *transparency* is widely cited by all countries around the world, to which can be added *confidentiality* in North America, *fairness* and *security* in Europe and *accountability* in Asia. In another study, Roche et al. (2022) looks for key terms in the documents to identify missing themes and show the under-representation of populations from the global south.

To go deeper in the understanding of these principles, Ryan and Stahl (2020) provides a more detailed description of the content that they cover, unpacking them into more refined themes. Yet, simply looking at the principles does not prevent the same word from having different meanings from a country to another. For instance, *privacy*, or *fairness* may be understood differently in the EU or in China (Fung and Etienne, 2022). To address this, one can look at the text as a whole, beyond the principles, and see if the vocabulary used differs by sector or country of origin. To investigate these differentials, (Rees and Müller, 2022) studies the frequency of identified keywords. For instance, Google or the UK government widely mention *"bias"* and *"fairness"*, but not *"diversity"* unlike the European Commission. Probasco et al. (2023) and Toney-Wails et al. (2024) also uses frequency analysis, to look for principles related to "trustworthy AI" in research articles, based on a list of keywords from the US National Institute of Standards and Technology. They find that policy documents and research articles tend to align on the different principles that are used (Toney-Wails et al., 2024). However, a closer analysis of the concrete definitions given for the same principle shows that different stakeholders use the same term to talk about different things. For instance, while some charters present "fairness" as a major problem of AI systems, others present AI as the solution to the problem of unfairness created by human biases (Evers, 2024). These nuances suggest that simply looking at the occurrence of the word "fairness" is not sufficient to understand the content of AI ethics documents.

In conclusion, previous meta-analyses are often centered around the study of principles, but few of them look at the text as a whole and, to the best of our knowledge, none has applied text analysis to AI ethics charters. However, such approaches have been applied to other types of documents. In relation to the ethics of AI, text analysis has been used to analyse documents related to sustainable AI in energy (Saheb et al., 2022), engineering ethics education (Nasir et al., 2021) or even national AI strategies (Papadopoulos and Charalabidis, 2020; Papyshev and

[9] The *humanity* principle defined in Zeng et al. (2018) encompasses, amongst other things, *human rights*, *dignity*, *freedom* and *well-being*.

Yarime, 2023). With this work, we intend to fill this gap and apply these methods to AI ethics documents.

2.2.3 *Shortcoming of ethics charters*

The various initiatives to guide AI ethics have been lauded around the world and has contributed to opening up the dialogue between different stakeholders on AI benefits and risks. They can also provide tools to measure the ethical outcome of a decision and help rethink professional practices and societal life (Langlois and Régis, 2021). According to Langlois and Régis (2021), they are a stepping stone to developing AI regulation and binding norms. However, they are also widely criticised.

Representation and meanings

First of all, the field of AI ethics and its charters suffer from a black box problem, even though the charters themselves vigorously criticise the opacity of AI systems (Bélisle-Pipon et al., 2022). Indeed, it is not always clear who develops these ethical guidelines and how: Bélisle-Pipon et al. (2022) found that only 38% these documents report stakeholder engagement, and 15% explain their methodology for arriving at their recommendations. Moreover, most of those charters are elaborated by and for western countries (Goffi and Momcilovic, 2022). As such, some geographic areas such as Africa, South and Central America and Central Asia are systematically underrepresented in the production of these documents (Jobin et al., 2019). Yet, they often claim to be universal, even though they contain specific cultural values.

Even where common objectives are identified, there are many differences in the interpretation of ethical principles, their presumed importance, the areas or actors to which they apply and their implementation (Jobin et al., 2019). The fuzzyness of ethical principles simplifies complex ethical debates and gives the impression of a consensus (Ryan and Stahl, 2020). In the words of (Munn, 2022): *"Who could be against beneficence? However, problems immediately arise when we start to define what beneficence means."* Indeed, simply looking at the principles does not prevent polysemy, accross countries and contexts. For instance, *privacy*, or *fairness* may be understood differently in the EU or in China (Fung and Etienne, 2022). To address this, one can look at the text as a whole, beyond the principles, and see if the vocabulary used differs by sector or country of origin. To investigate these differentials, (Rees and Müller, 2022) studies the frequency of identified keywords. For instance, Google or the UK government widely mention *"bias"* and *"fairness"*, but not *"diversity"* unlike the European Commission. However, (Rees and Müller, 2022) defines the keywords manually and only displays results by document, not by sector or country. To our knowledge, no temporal analysis of these documents, to see if certain principles are mostly cited in older texts and if some have emerged in recent ones, has been done yet.

Difficult implementation

Another widespread criticism concerns the gap between the principles set out in these texts and their concrete implementation (Tidjon and Khomh, 2022). Indeed, AI ethics charters never explain how to apply these criteria in real-life contexts, contenting themselves with describing perfect systems, independently of the sociological, economic and technical conditions surrounding them. The vagueness of their recommendations leads industrial stakeholders to dismiss them, either not knowing how to proceed to respect them, or claiming to comply with them without changing their processes. Indeed, many of the guidelines set out broad principles, as well as ethical and social issues that are not even specific to AI (Hagendorff, 2022). As a result, machine learning practitioners tend to find them of little use (Henriksen et al., 2021). Even when they are instructed to take ethical guidelines into account, the change is not significant enough to alter the way they work (McNamara et al., 2018) and while they acknowledge the importance of ethics, practitioners do not take it into account in their development practices (Vakkuri et al., 2019).

As these texts are usually heavily focused on AI products or services and do not take into account the general context in which such systems are developed, they thus fail to capture the underlying social issues. For instance, the "biases" of AI algorithms are not solely technical issues but can be traced back to the sexism, misogyny and racism latent in the tech industry and society at large (Munn, 2022). Any guidelines dismissing this context is thus bound to be ineffective.

It is also common for these principles to be presented orthogonally, as independent guidelines, hiding the fact that they are always in tension with each other and cannot all be all satisfied at the same time, or at least not to the same degree. Conflicting normative values require trade-offs, which imply a hierarchy of values: the use of personal data that would improve public services undermines privacy (Morley et al., 2021), to give only one example. This hierarchy is never discussed, but this does not mean it does not exist; on the contrary, it is considered universal and not worthy of interest[10].

Ethics-washing and performativity

On the contrary (Hagendorff, 2020) argue that, far from being inoperable, AI ethics guidelines are today limited to what is measurable. As a result, these tools which are supposed to promote better systems become simple checklists for designers to follow, not using their ethical reasoning skills, but ticking boxes on a list of mere good practices (Balayn et al., 2023). As a result, they miss out on the bigger issues of AI technologies, such as the harm caused to marginalised communities (Kalluri, 2020).

Ethical initiatives are often over-written by economic incentives and are simply used for marketing purposes and communication strategies (Munn, 2022). The use of ethics by private companies can also be a strategy to avoid stronger regulation (Nemitz, 2018). Ethics is thus used as a cover for unethical behaviour, either because the principles are too vague and therefore considered by companies to have already been implemented, or because the list of things to check is not very restrictive and already corresponds to what is usually done in the field. Scholars refer to this trend as "ethics-washing" (Wagner, 2018; Bietti, 2020). As a result, ethics guidelines are seen, at best, as useless (Munn, 2022) and, at worse, as a tool used by lobbyists for regulatory capture (Saltelli et al., 2022).

Furthermore, these guidelines give little advice on who should be responsible for ensuring the ethical requirements they comprise. There is a lack of legal and professional accountability mechanisms to AI ethics, unlike in medical ethics

[10] See (Birhane et al., 2022) for a discussion of the values encoded in machine learning research.

for instance (Mittelstadt, 2019). Guidelines are not binding nor is there an entity that controls their enforcement. Today, these principle-based guidelines *"lack teeth"* (Rességuier and Rodrigues, 2020).

There is a growing set of tools – checklists, standards, risk assessments, open-source libraries and so on – that is being developed to address AI ethics. However, Morley et al. (2020) argues that they are difficult to map with regards to the principles they could help to address. For Ryan and Stahl (2020), this mapping between ethical guidelines and technical tools is the missing step towards the adoption of these frameworks in practice.

2.3 *Corpus collection and overview*

We now detail our first contribution, the curation of a corpus of documents discussing "AI ethics". We detail our collection process, the formatting of the data, the preprocessing that was uniformly applied to the whole corpus, and finally its availability and ways of future contribution.

2.3.1 *Collection*

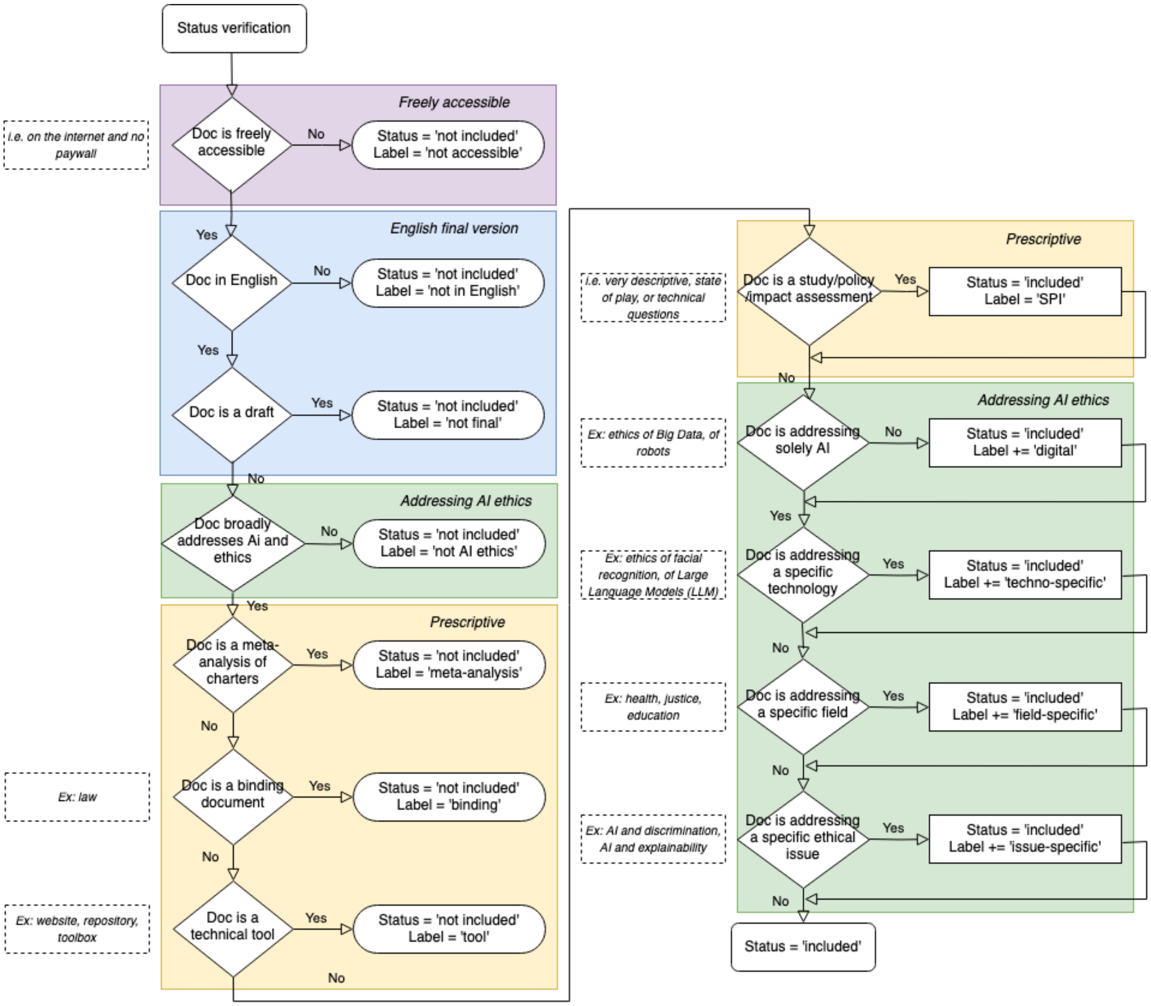
To choose which documents to collect, we referred to several existing repositories and meta-analyses. Table 2.1 shows the overlap between our corpus and previous works, showing that our database is the second-largest, behind the one compiled by the Council of Europe. We obtain a list of documents that were cited at least once in one of the previous works. Since our goal is to provide a quantitative outsight on previous papers, we refrained from adding documents that have never been considered in previous studies, though they do exist. In total, we annotated 730 documents and filtered them using the following list of inclusion criteria:

1. The document must be freely accessible: we discard any document that we cannot find, that is behind a paywall, or that requires subscription to access;
2. The document must be written in English, and not be in a draft state: we do not consider documents in another language, or unofficial translations;
3. The document must discuss artificial intelligence and AI ethics;
4. The document must be prescriptive: we do not include binding documents, standards, purely technical documents, or any purely descriptive documents. In the case of a largely descriptive document with a few prescriptive recommendations, we include the document and label it "SPI" (Study, Policy or Impact assessment).

We summarise our process as well as the number of documents filtered out at each step in Figure 2.2. Our rationale for selecting documents is guided by the desire to have a quality analysis of the documents. Having documents of the same nature allows for a more relevant comparison of the vocabulary used. This guides each of our inclusion criteria.

First, we remove non accessible documents. Not accessible might refer to paywalled documents, not found documents, or documents that we cannot automatically scrape (for example, multiple web pages).

We exclude non-official translations to avoid misunderstandings when we cannot ensure the quality of the translation, or when the translation itself imposes an



(a) Flowchart documenting our annotation process. Each rhombus corresponds to a choice (of inclusion or annotation). The left column corresponds strictly to exclusion criteria, such as language, accessibility, prescriptivity, while the right column all consider included documents with specific, cumulative comments, such as documents focused on specific fields (e.g. healthcare).

Step	# of documents
Considered documents	730
Accessible	677
In English, is final	608
Addresses AI and AI ethics	510
Is prescriptive	436
Included	436

(b) Table summarizing the number of documents discarded at each step. Reading key: $730 - 677 = 53$ documents were discarded because we could not access them.

Figure 2.2: Flowchart and numeric breakdown of the inclusion criteria for the collection of the MapAIE corpus.

unchecked western bias. For instance, in the document titled "Advisory Board on Artificial Intelligence and Human Society", an initiative of the "Minister of State for Science and Technology Policy" included in (Jobin et al., 2019), the Chinese term usually translated as "harmony" in English, which comes with moral and social preconceptions that are closer to the translator than the original intent; for a concrete example, the interested reader can read the work of Werbach on the Chinese social credit systems (Werbach, 2022). Keeping only the latest versions and official releases allows us to respect the authors' words and to discard obsolete statements.

Selecting only prescriptive documents permits us to discuss how AI should be. On the contrary, more binding documents usually restrain their scope to what is possible or desirable with other constraints (economic, social or technological ones) and thus rather discuss how AI could be. Similarly, study on the state of AI ethics in the world rather discuss how AI is today or will be in the future.

We apply our annotation process to the 730 potential documents, and we include 436 of them in our corpus, only including documents that have been cited in at least one previous meta-analysis. Each document was assigned for review to one of the authors, and so we have 4 annotators. To ensure consistency between annotators, we collegially annotated 10 documents, and then selected 10% of the original base to be blindly annotated a second time by three of the four annotators. We measure inter-annotator agreement with Fleiss's κ, which takes its values between -1 (perfect disagreement) and 1 (perfect agreement), a value of 0 indicating a chance assignment. We obtain $\kappa = .712$ $(95\% CI, .577$ to $.847, p < .001)$, indicating high agreement between the annotators[11]. We break down in Table 2.1 the overlap in included documents with previous papers. It shows that we included documents used in a variety of studies. However, we could not include all of them, as many did not meet our inclusion criteria.

		MapAIE	Jobin et al.	Fjeld et al.	Tidjon and Khomh	Hagendorff	Floridi and Cowls	Zeng et al.	Attard-Frost et al.	European Parliament	Algorithm Watch	Council of Europe	Winfield	EthicalML
MapAIE (our paper)	(100.00%)	436	73	32	15	20	6	65	35	9	114	360	20	12
Jobin et al. (2019)	(87.95%)	73	83	20	3	18	6	32	22	7	74	69	18	6
Fjeld et al. (2020)	(86.49%)	32	20	37	2	12	5	21	31	3	27	31	11	6
Tidjon and Khomh (2022)	(51.72%)	15	3	2	29	3	2	4	3	2	4	8	2	5
Hagendorff (2020)	(95.24%)	20	18	12	3	21	4	14	16	3	18	17	6	4
Floridi and Cowls (2019)	(100.00%)	6	6	5	2	4	6	6	5	2	6	6	4	2
Zeng et al. (2018)	(78.31%)	65	32	21	4	14	6	83	24	6	50	54	13	6
Attard-Frost et al. (2022)	(76.09%)	35	22	31	3	16	5	24	46	4	32	36	10	7
European Parliament (2020a)	(75.00%)	9	7	3	2	3	2	6	4	12	7	7	6	1
Algorithm Watch (n.d.)	(71.70%)	114	74	27	4	18	6	50	32	7	159	122	19	8
Council of Europe (2022)	(60.50%)	360	69	31	8	17	6	54	36	7	122	595	18	10
Winfield (2019a)	(83.33%)	20	18	11	2	6	4	13	10	6	19	18	24	4
EthicalML (n.d.)	(80.00%)	12	6	6	5	4	2	6	7	1	8	10	4	15

Table 2.1: The matrix of documents in our dataset (MapAIE), compared to previous works. Reading key: 87.95% of documents in Jobin et al.

[11] https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

2.3.2 *Formatting*

We list all documents, their title and institution of origin, the URL address at which we reach them, and our annotations in a tabular file. All documents are either in PDF or HTML format. We automatically download each document, and extract its contents using Python scripts. In the case of PDF files, we use the Python library `PyPDF2`[12]. In the case of HTML files, the situation is more complex, as just downloading the page includes a lot of boilerplate content (menus, headers, links to other pages, etc.). We design an algorithm to extract the main content of the page, by finding the deepest element in the HTML structure tree (DOM) that contains the largest content.

2.3.3 *Pre-processing*

We automatise preprocessing for text fields. All text is processed using the python libraries `BeautifulSoup`[13] and `NLTK`[14]. `BeautifulSoup` is designed to manipulate HTML structures and extract textual contents; the *Natural Language Toolkit* (`NLTK`) provides tools for working with human language data, for the text itself. We systematically remove numbers, URLs, and stop words[15] present in the NLTK `english stopwords` corpus, and put all text in lowercase. Then, we retrieve all the lemmas appearing in the text, i.e all the canonical forms corresponding to the words[16] composing the text; for example, the lemma "train" corresponds both to the words "training" and "trained". Finally, we remove all lemmas that contain less than 3 characters.

Our final corpus comprises of 436 documents. We release online[17] the tabular file listing all documents (included or not), the corpus itself, as well as its datasheet (Gebru et al., 2021) and the parsing and preprocessing code. Due to intellectual property limitations, we cannot publicly release the scraped contents as is. Instead, we release the code required to download and build the corpus in a single command. All materials are available publicly, on academic storage (provided by our institution), as well as on GitHub. In order to ensure reproducibility and open the way to new analyses, we publicly document our process, allowing individuals to include new documents so that anyone can contribute to enlarging the corpus, provided they follow our annotation guidelines.

2.3.4 *Creating thematic corpora*

From the initial corpus, we build several thematic corpora along guiding themes identified in previous meta-analyses. These corpora do not form a partition of the corpus: a document can belong to multiple corpora. We specifically discuss analysis and results of these subcorpora along themes we identified and themes identified by Jobin et al. (2019). We display the themes we identify and the associated keywords in Table 2.2, and show the co-occurrences of themes in Table 2.3.

2.4 *Analysing themes in the corpus*

2.4.1 *Exploratory analysis*

Let us start by examining a few generalities about the corpus. First of all, a comment on the length of the documents. We nuance the common preconception

[12] `https://pypi.org/project/PyPDF2/`

[13] `https://pypi.org/project/beautifulsoup4/`

[14] `https://www.nltk.org/`

[15] Words that are very commonly used in a language, such as "the", "is", etc. in English.

[16] We recognise that the term "word" is not the one generally used in linguistics to describe a textual content. For the sake of simplicity, we use it in this article to stand for "token" or "word form", or "lexeme".

[17] `http://mapaie.telecom-paris.fr`

Theme	Keywords
fairness	fairness, algorithmic fairness, bias
xai	xai, lime, shap
regulation	personal, right, law, harm, gdpr, discrimination, article, biometric, regulation
agi	agi, artificial general intelligence

Table 2.2: Our themes and keywords.

	Fairness	XAI	Regulation	AGI
Fairness	0.48	0.04	0.37	0.05
XAI	0.04	0.04	0.03	0.01
Regulation	0.37	0.03	0.51	0.05
AGI	0.05	0.01	0.05	0.07

Table 2.3: Co-occurrences of themes in our corpus.

that AI ethics charters are short and of little practical use (Henriksen et al., 2021): we notice instead a difference between purely positional statements and more fleshed-out documents, with roughly 20% of documents exceeding 10000 words (around 20 pages of text).

Word	Frequency
data	43412
systems	16852
use	16663
intelligence	16242
artificial	14702
human	14334
also	13583
public	12126
rights	11759
system	11757
research	11485
may	11234
development	10195
digital	10186
new	9907

(a) Lemmas with the highest term frequency across all documents with their total word counts.

Word	# Documents
use	331
data	331
public	319
information	318
development	318
also	317
intelligence	314
human	314
systems	313
new	311
research	306
make	305
society	305
privacy	305
social	304

(b) Lemmas with the highest document frequency with their total document occurrence.

Words	Frequency
artificial intelligence	43669
data protection	39668
personal data	36742
machine learning	34862
human rights	33757
ai system	33031
data use	31289
data protection regulation	30926
data collection	30179
public sector	30149
european commission	30130
impact assessment	29891
member states	29118
general data protection regulation	28411
best practices	27333

(c) *n*-grams with the highest co-occurrence frequency.

Table 2.4: Term frequency and document frequency of lemmas in the corpus. Reading key: the lemma "data" appears 43412 times in the whole corpus; it is used in 331 documents among the 436 that constitute our corpus and is the second most used lemma. It occurs in a bigram with the lemma "protection" 39668 times, in a trigram with the lemmas "protection regulation", and in a 4-gram.

Most frequent terms across the corpus are represented in Table 2.4a. Terms like "system" and "data" are over-represented, while other lemmas follow a rapid decay. Notably, "artificial intelligence" is much less used than the term data for instance. Yet we need to keep in mind that the lemma "AI" is removed during preprocessing because it is too short and thus does not appear in this list. Document frequency, however, follows a much slower decay (Table 2.4b): many terms are present in several documents. The term most common to documents is "use", followed by "data" and "public". "Artificial intelligence" only appears in 314 documents out of 436; the remaining 122 documents typically discuss AI in a narrower sense, e.g. "machine learning for face recognition", or use the word "AI" without explaining what it stands for, which we deemed fully in scope.

We show in Table 2.4c the most frequent *n*-grams, i.e. sequences of *n* words that frequently co-occur together (for example, "artificial intelligence" is a 2-gram, and "data protection regulation" is a 3-gram). *n*-grams give us more meaningful insights into the themes and discussions of the “artificial intelligence”, “data protection”, “machine learning” and “human rights” come up as very frequent, with most of the top *n*-grams being related to legal and regulatory texts (“personal data”, “fundamental rights”, etc.). It also highlights the central role of European institutions as regulators of artificial intelligence as of the writing of this paper.

2.4.2 *Understanding recurring themes and common topics*

We continue our study by an analysis of common themes in our corpus. We first analyse the whole corpus in Figure 2.3, and discuss the main currents of thought we find. The clusters are built using the Reinert method (Reinert, 1990), a hierarchical clustering method, and the results are visualised with a correspondence analysis (Hirschfeld, 1935). Each text in the corpus is analysed through the lens of co-occurrences of lemmas in fixed size text segments. We use segments of size 40, though we examined different segment sizes (between 2 and 200) to ensure the stability of the results. All analyses were made using the IRaMuTeQ software[18].

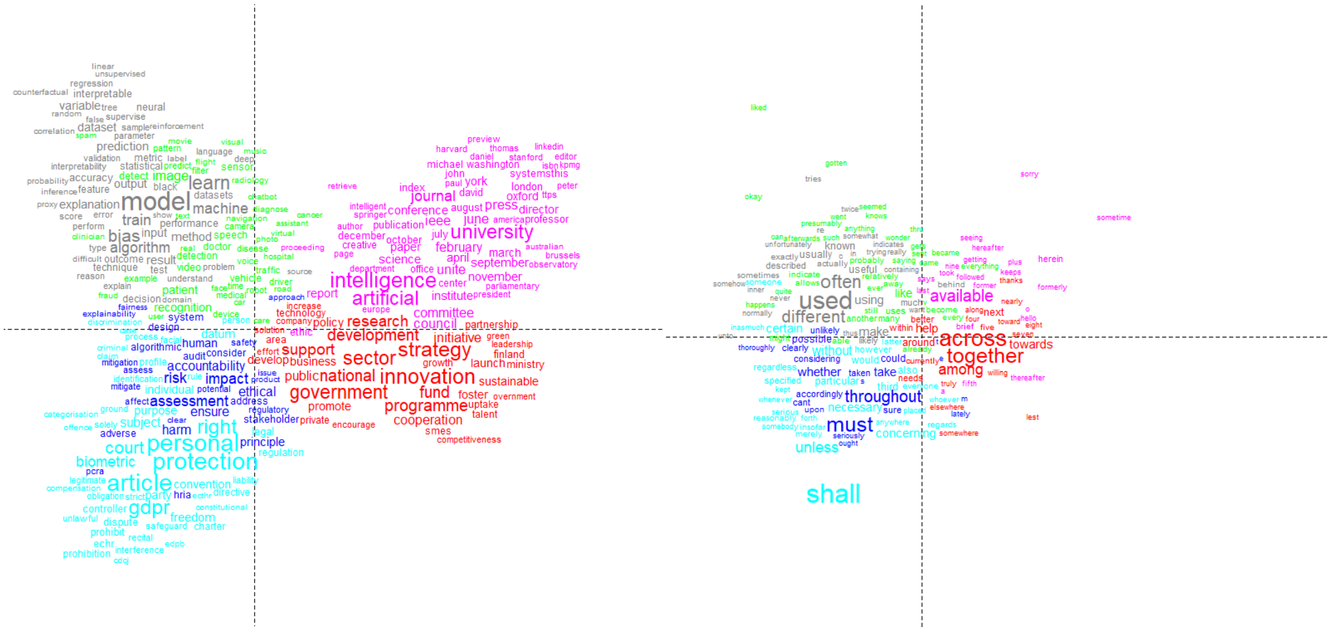


Figure 2.3: Two-dimensional visualisation of the clusters obtained with hierarchical classification on our corpus, obtained with correspondence analysis. The size of words is proportional to their importance (in terms of number of occurrences) in the corpus, and distances are linear. Explained variance: 61.5%.

The general clustering in Figure 2.3 highlights 6 clusters, corresponding to different themes: two of them are technical (centered around models/techniques and applications, respectively), two are more regulatory (centered around laws and policies, respectively), and the last two correspond to a business-oriented and a very generic cluster, respectively. In Figure 2.3 right, we see how different common words are associated with each cluster: while the technical and applicative clusters use descriptive language (“used”, “often”...), the regulatory cluster uses prescriptive one (“must”, “shall”...).

To each main current (technical, regulatory, innovation) corresponds a different paradigm: the technical documents largely follow a model-driven paradigm, while the regulations and laws follow a data-driven paradigm; finally, documents discussing innovation largely frame it as strategies, programs and plans in order to keep a competitive edge. We note the absence of a “user-driven” paradigm, examining the role of human beings in relation to AI and its ethics. Though this is

[18] http://www.iramuteq.org/

partly captured by regulation and law in the form of "data", the correspondence between human and (personal) data is nothing but systematic, even though it is a common assumption of machine learning models (Cardon et al., 2018). Indeed, unlike humans, data is typically reduced to atoms of information and vector-based. This irreducibility, along with works around the ethnography of algorithms, studying how end-users react and use data algorithms, have shown effects of resistance and decoupling between institutional discourses and practical use (Christin, 2017, 2020). We also note the absence (at least at this scale) of strong discussions on social justice issues, even though sexism, racism (Phelps, 1972; Zou and Schiebinger, 2018) and labour inequality (Tubaro et al., 2020) are well-documented problems in artificial intelligence models and datasets.

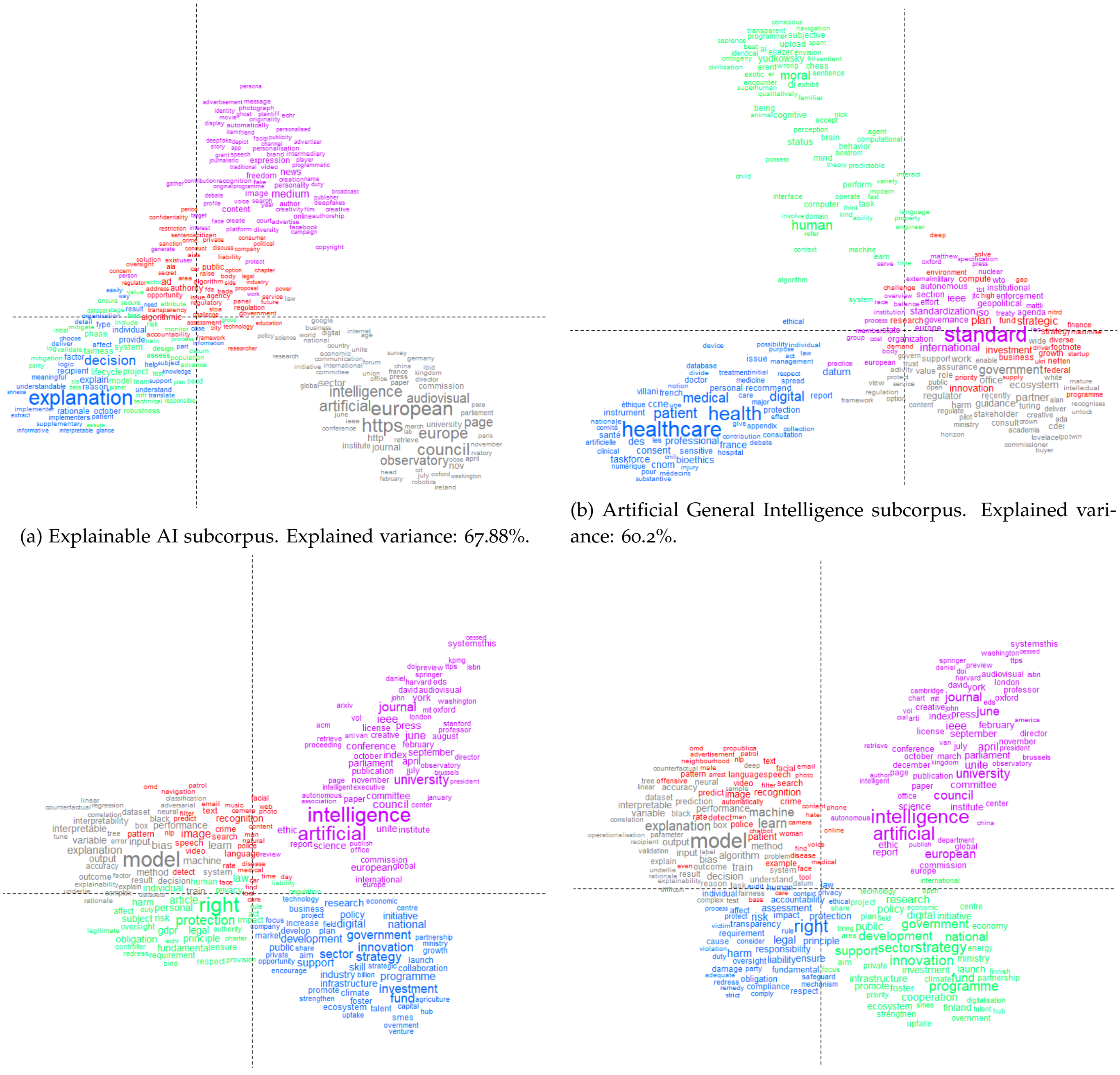
Analysing themes in the corpus

We analyse themes that follow (Jobin et al., 2019) in Section 2.4.2, but we have also decided to expand this analysis to themes that have emerged since 2018. Our hypothesis is that analysing these themes brings a complementary perspective. The visualisations of these analyses are presented in Figure 2.4, each subfigure corresponding to one of the themes outlined in Section 2.3.4.

We notice that *explainable AI* (XAI, Figure 2.4a) remains a technic-dominated area, with very specific technical vocabulary (explanation, decision; bottom-left cluster), with another technical cluster on top-right more centered around applications of explainable AI, with the terms "deepfakes", "content", "diversity", "fake", etc. Well separated is a regulation cluster (bottom right), centered around the European Union, with few meaningful words. In the case of the *Artificial General Intelligence* (AGI, Figure 2.4b), a term that is commonly tied to the moral panic that AI systems will overcome human beings in the long-term, we see that the technical cluster completely disappears, while the regulatory one drastically shrinks: in other words, AGI is not a topic of interest from the technical point of view, and marginally so in the case of regulation. Instead, the terms mobilised focus on standardisation (bottom right), human and moral considerations (top), and medical and health considerations. Quite interestingly, the last two subcorpora, related to *fairness* (Figure 2.4c) and *regulation* (Figure 2.4d) are both similar to the global analysis of the corpus. We take away from this that (i) fairness has become a commonplace term, that is reproduced in all areas of "AI ethics" (though, possibly with polysemy), and (ii) that most documents in our corpus discuss regulation, indirectly or not.

Confronting with themes in the literature

The analysis presented in the previous section gives us an opportunity to confront the corpus against recurring themes identified in the literature. We filter our corpus using the keywords outlined in (Jobin et al., 2019); we then run the same preprocessing, clustering and correspondence analysis on each sub-corpora. When using the keywords and themes identified in (Jobin et al., 2019), while the clusters' words change marginally, the gist of the results stay the same, with clusters separating the data along three lines: technical, regulatory and innovation/business, in addition to a more generic cluster. The reasons for this relatively small changes are multifactorial: firstly, the keywords listed in (Jobin et al., 2019) are quite generic



(a) Explainable AI subcorpus. Explained variance: 67.88%.

(b) Artificial General Intelligence subcorpus. Explained variance: 60.2%.

(c) Fairness subcorpus. Explained variance: 64.1%.

(d) Regulation subcorpus. Explained variance: 66.6%.

Figure 2.4: Thematic analysis along our subcorpora, visualised with a correspondence analysis.

(listing, among others, "disclosure" and "showing" under the theme *Transparency*). This is not necessarily a problem in the original case, which focused on a qualitative analysis and where researchers can decide on a case-by-case basis if a word matches a theme; furthermore, the concepts have spread across actors and institutions since 2018, year of (Jobin et al., 2019), and they are now sufficiently widespread that they are not markers of differenciation anymore. We list here the themes the authors identified, explaining the key changes they induce in terms of text analysis, i.e. how the four main clusters (technical, legal/regulation, innovation and generic) evolve and change; a cluster becoming smaller and more specific is typically due to less documents discussing this paradigm in the subcorpus.

Transparency (257 documents). There are no changes along this theme, showing how transparency has permeated discourses around AI ethics and is now used indiscriminately in technical, regulation and innovation documents.

Justice and fairness (78 documents). Another widely used theme. The business and innovation cluster shrinks in terms of size, while the legal and regulation one becomes larger; the technical cluster becomes more specific, explicitely citing algorithmics fairness related terms.

Beneficence, non-maleficence (83 documents). In this case, the technical and regulation clusters get closer and tighter, while the innovation and generic clusters remain mostly unchanged. This is due to both technical and regulation documents mentioning these topics, in extremely similar terms.

Responsibility (154 documents). There are no cluster changes along this theme, even though the technical cluster shrinks in size, and become slightly more specific.

Privacy (106 documents). The legal and technical clusters fuse into a single one, highlighting more specific applications (such as, for example, "homomorphic cryptography", leaving the rest relatively unchanged.

Freedom and autonomy (25 documents). In this theme, the clusters become more specific, discussing jobs and work-related issues, specific technical terms such as "bias" or "model manipulation"; other clusters gather terms related to creativity and cooperation, along with a small regulation cluster focused on the implementation of legal texts.

Trust (279 documents). There are no specific changes, showing that the topic has permeated AI ethics.

Sustainability (159 documents). While the core results remain unchanged, the law/regulation and technical more separated, indicating less overlap in how these topics are discussed by regulatory and technical documents.

Dignity (124 documents). The main results do not change, apart from the legal and regulation cluster becoming much larger than the technical one. Indeed, *dignity* has a strong legal connotation and is routinely used in this context.

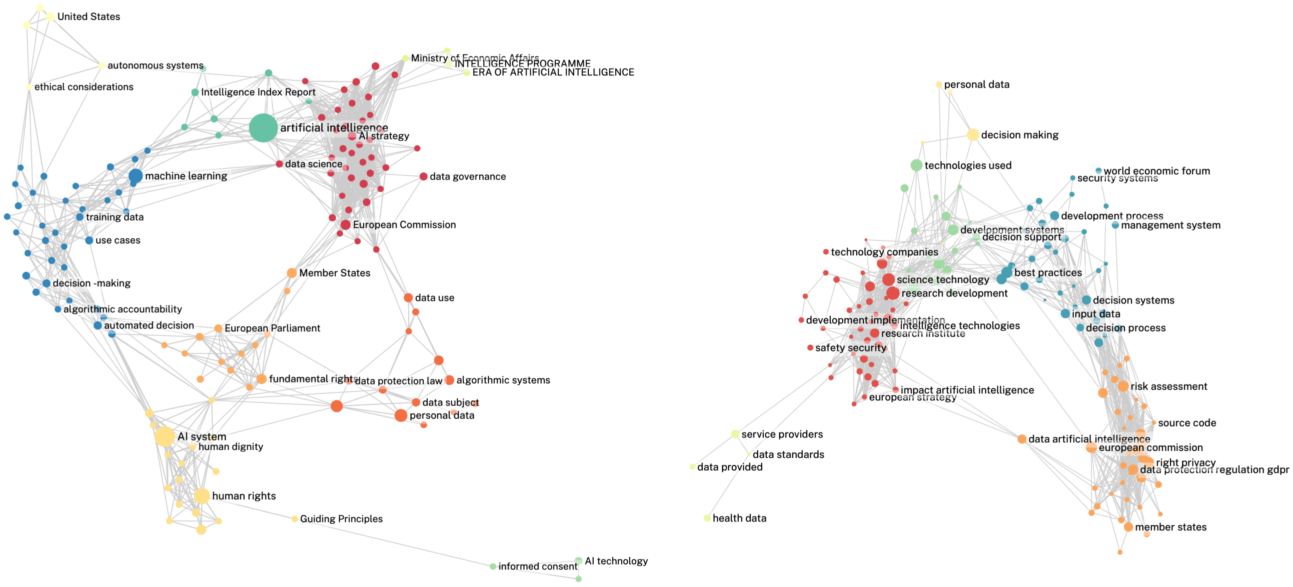
Solidarity (32 documents). The legal/regulation and innovation clusters remain stable. However, the technical cluster becomes more specific (citing terms around "interpretability", "explanation", "fairness"...), and the generic cluster is replaced by a more interesting one, centered around jobs, employment and economy.

In conclusion, while some themes have been consistently picked-up on by the various actors and institution, this is not the case for all of them, especially the more specific ones. The number of documents associated to each theme sorts the themes in a different order than the one in (Jobin et al., 2019), though we are not the first to notice this (see Section 2.2).

2.5 *Areas of consensus and of confrontation*

Delving into the specific vocabulary, and relative importance of the areas around which discourses are structured, this gives us the possibility to look into both consensual and confrontational areas. In this section, we use semantic graphs to identify some controversies inherent to modern artificial intelligence. These graphs, by showing us words that are at the frontier of clusters (i.e. typically linked to nodes of their own cluster as well as other ones, as for example "Member States" in Figure 2.5a), show us the themes where semantic and semiotic qualms happen.

2.5.1 *Building co-occurrences graphs*



(a) Co-occurrence graph of *n*-grams.

(b) Co-occurrence graph of *n*-grams filtered using TF-IDF.

Figure 2.5: Co-occurrence graphs. Reading key: the node corresponding to the bigram "artificial intelligence" is part of the green cluster in the left graph. It connects the blue cluster with the red cluster. It is represented with a wide circle as it is highly connected.

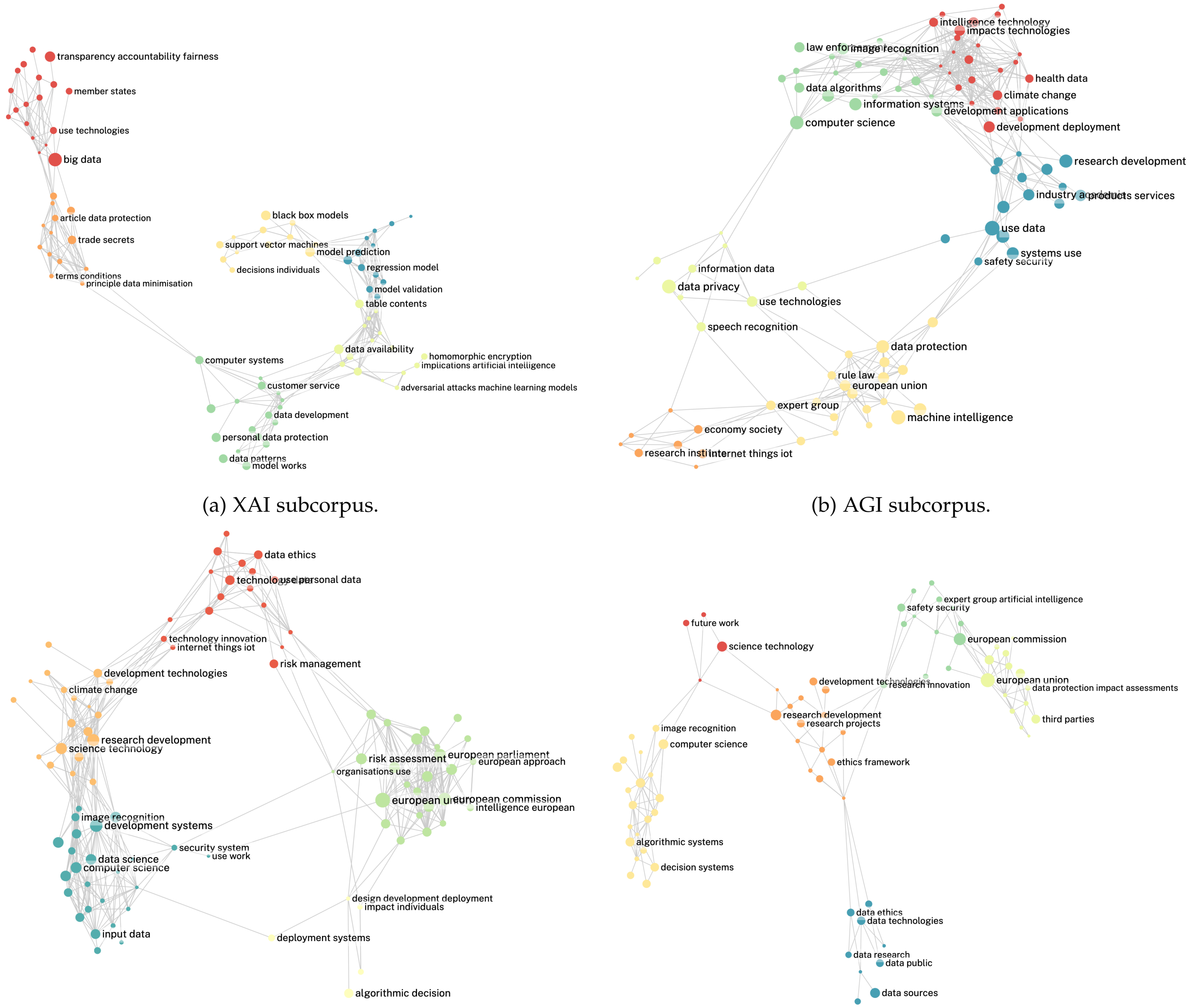
We build co-occurrences graphs. A graph is a tuple $G = (V, E)$, where V is a set of nodes $(\{u, v, w, \dots\})$ and E a set of edges (i.e. pairs of nodes, $\{(u, v), (u, w), \dots\}$). We will consider graphs to be undirected (i.e. $(u, v) = (v, u)$) and loopless (i.e. $u \neq v$). We build the graphs so that nodes are n-grams in the corpus (with $2 \leq n \leq 5$), and there is an edge between two nodes if the n-grams significantly co-occur in the corpus. Significance is tested via a chi-square (χ^2) test, which compares the observed and expected frequencies of the outcomes of variables. The size of the node in the visualisation is propotional to its degree, i.e. the number of connections with other nodes: the higher the number of connections, the bigger the circle representing the node. Nodes are then colour-coded using the Louvain algorithm (Blondel et al., 2008), a common graph clustering algorithm that detects subsets of nodes that are more connected together than with the rest of the graph, by optimising an objective function. Notice that, due to its aggregative design, Louvain typically favours larger clusters.

2.5.2 *Analysis of co-occurrences graphs*

We show in Figure 2.5 graphs built from our corpus[19]. The left co-occurrences

[19] Interactive graphs are available at (a) and (b).

graph displays relationship between *n*-grams. The right co-occurences graph displays the same relationship with *n*-grams filtered on important words, as per the *Term Frequency – Inverse Document Frequency* (TF-IDF) metric, which measures the extent to which a word appears a lot in a document (*Term Frequency*), but seldom in most documents of the corpus (*Inverse Document Frequency*).



(a) XAI subcorpus.

(b) AGI subcorpus.

(c) Fairness subcorpus.

(d) Law subcorpus.

Figure 2.6: Thematic graph analysis, along our subcorpora. A link between two terms means that they co-occur significantly in the subcorpus.

From Figure 2.5a, we notice the predominant position of the term "artifical intelligence", connecting two major communities related to AI techniques (blue) and governance (red). We observe that "machine learning" belongs to technical usage, while business actors and impact assessment writers tend to focus more on "data science". Interestingly, while the European Parliament and Council are together in a cluster related to fundamental rights (orange), they are separated from the European Commission, which is closer to governance topics (red). This outlines the role of the European Commission as a provider of expertise, rather than a regulatory or legislative instance.

Filtering *n*-grams on the most important terms allows to avoid the influence of generic terms such as “artifical intelligence”. In Figure 2.5b, clusters are slightly modified and we find four major communities: (i) Research & Development (red); (ii) technological systems (green); (iii) management and process (blue); and (iv) protection and regulation (orange). Interestingly, the importance of individual rights related nodes is lowered after considering TF-IDF; “human rights” or “human dignity” disappear to the benefits of themes such as “right privacy” or “data protection”. Moreover, we notice the absence of terms such as “fairness”, “ethics” or “explainability”, as they appear widely through the corpus: the terms “ethic[s | al]” appears in 81.6%, “fair[ness]” in 72.05%, “explain[able | ability | ation]” in 67.4% of documents. Overall, we observe strong semantic proximity between technically-oriented clusters (red and green), but highlight how distant such considerations can remain from operational and economical aspects (blue) as well as from regulation vocabulary (orange).

We further analyse thematic co-occurrences graphs by filtering our corpus using the keywords in (Jobin et al., 2019)[20]. We observe in Figure 2.6c that filtering the data using commonly used terms such as “Fairness” only induces minor change in the co-occurrences graph; the different clusters and their relationships remain stable. Similarly, the “Artificial General Intelligence” (AGI) graph, in which all communities are kept in their original proportions, suggests that the term is broadly used by all categories of actors in the AI world. On the other hand, by focusing on documents containing “XAI”, we exhibit a highly technical graph where regulatory considerations are almost not present at all. At the other side of the spectrum, the “Regulation” graph in Figure 2.6d evokes several aspects of AI regulation in addition to the technical references. However, we observe how business oriented terms are absent from this perspective. These two examples suggest a strong semantic boundary between these two worlds.

2.6 Analysing documents from European institutions

2.6.1 Selecting documents

Examining documents from around the world allows us to understand the different influences on AI ethics discourses. This thesis, however, discusses AI regulation in Europe specifically. We have therefore decided to look specifically at documents produced by European institutions – the Commission, the Council and the Parliament (61 documents). We have computed the clusters using the same method as described in Section 2.4.2.

2.6.2 Cluster analysis

The clustering method used on these EU documents highlights 4 clusters, shown in Figure 2.7, which do not correspond to what we found on the total corpus. The technical clusters from the total corpus have fused with the generic cluster (new “technical cluster”). The business cluster has became more about political strategy than business strategy, focused on national initiatives and funding programmes (“strategy cluster”). The regulatory clusters from the total corpus have

[20] Interactive thematic subgraphs are available at (a), (b), (c) and (d).

also fused into one, more oriented towards good policy than individual rights and legal frameworks ("regulatory cluster"). We also witness a new cluster, clearly oriented towards biometric technology ("biometric cluster"). The presence of the biometric cluster reveals that this type of technology, which includes facial recognition, is a central theme of EU policy documents on AI, even though only 2 out of 61 documents are devoted to it. On the contrary, other themes which also have their own documents, such as AI in health, justice, education or public services, do not have their own cluster. Contrary to biometrics and facial recognition, they are only discussed in these specific documents but do not permeate all EU documents on AI.

In Figure 2.7 right, we also see that the regulatory cluster and the biometrics cluster both use prescriptive common words such as "must" and "shall". The word "anywhere" in the biometrics cluster shows that EU institutions are aware of the widespread use of these technologies.

These results show that EU documents on AI are more focused on application and implementation than on regulatory incentives.

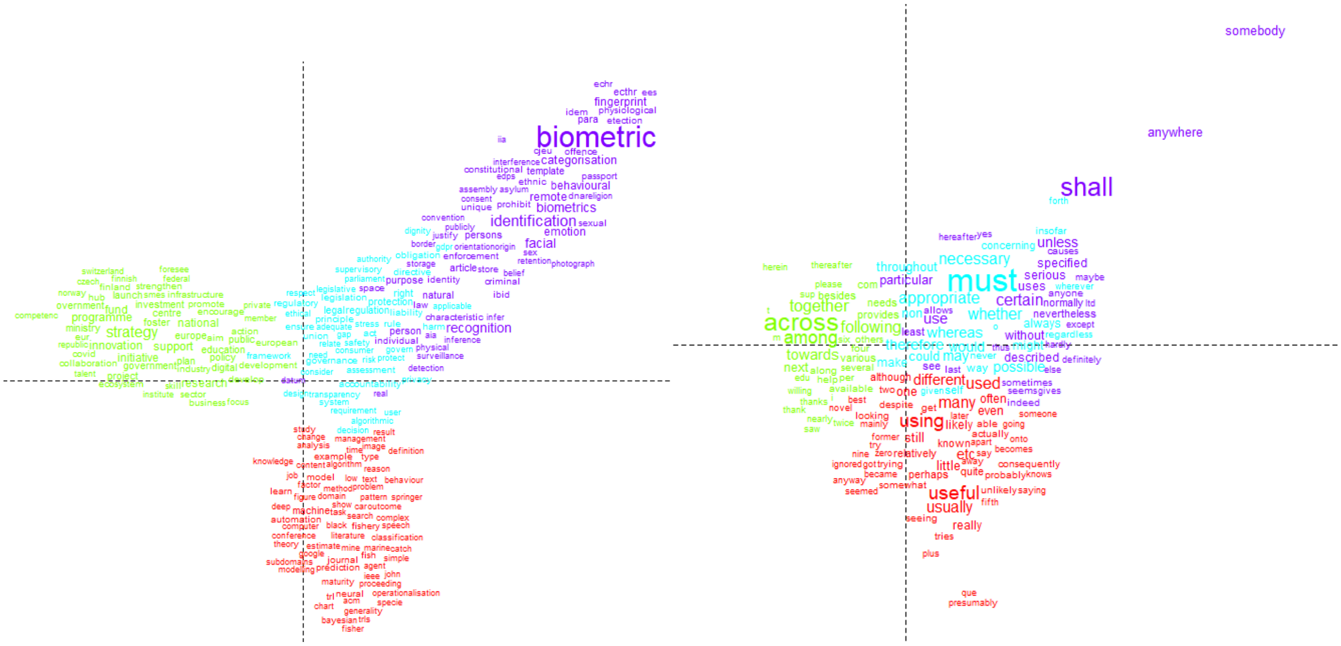


Figure 2.7: Two-dimensional visualisation of the clusters obtained with hierarchical classification on the corpus of EU documents, obtained with correspondence analysis. On the left, the strategy cluster; in the middle, the regulatory cluster; on the right, the biometric cluster; and in the bottom, the technical cluster.

2.7 *Limitations*

Let us outline some limitations of our work. The most obvious limitation is related to restraining our search to documents in English. Indeed, we made this choice to be able to compare texts on the same semantic level; but it leaves out multiple documents that have been written in other languages. We refrained from making any conclusions about the geographical origin of documents discussing AI ethics, even though we collected the data: we do know that our corpus is heavily biased in that regard. This bias stems notably from our country of origin, the language inclusion criteria, and the fact that we prioritised documents that were already mentioned in previous meta-analyses that exhibit such bias themselves.

Other limitations concern the methods used for our quantitative analysis. To begin with, the exploratory analysis is based entirely on word occurrences. However, this depends a lot on how the words are counted, which is influenced by our preprocessing method. For example, "AI" was filtered out by our preprocessing, so "artificial intelligence" has a lower word count than it would have if both versions of that term were counted together. Furthermore, both analysis methods we use are good at capturing common themes, rather than themes corresponding to less frequent terms or terms specific to one document. For instance, the theme of power struggles is not completely absent in the corpus but, because it is not statistically central, it is dismissed by the model. Lastly, for intellectual property reasons, we cannot publicly release the textual contents of the corpus, only make them downloadable. This means that documents becoming unavailable in the future will not be downloaded.

2.8 Discussion

2.8.1 AI ethics documents as a means of communication and influence

Conducting this study, we have read many AI ethics documents and learned a lot about what they can contain and we can therefore try to infer what the intentions of the writers were. This has allowed us to make high-level observations and identify general trends in our dataset.

One key takeaway is that, for the industry (and in particular AI BigTech companies[21]), academia and governmental authorities, providing a document discussing their vision of AI ethics acts a social signal (Connelly et al., 2011), proclaiming their willingness to participate in the discussion and their legitimacy to do so, regardless of the contents of the documents: the *existence* of such a document is as important as its content.

We also note a stark contrast between documents that describe how AI could be made ethical in the near future using already existing tools, and documents which design ethical requirements independently from the technical state-of-the-art. The distinction between "abstract" and "concrete" elements, also present in other analyses (Benbouzid et al., 2022), is of paramount importance, and addresses different audiences with different goals. Abstract communication typically focuses on high-level explanations of technical tools, that are introduced *as is* in regulatory texts. Indeed, industrial actors can afford to use precise technical language in the documents they write: they typically have a deeper understanding of the models, and their mathematical limitations. Institutional actors, on the other hand, have a more precise grasp of the social and legal implications.

In our results, we highlighted clusters of terms that most frequently appear together in documents. We argue that this separation in clusters shows that different sectors position themselves with different roles, and that the communication of abstract and concrete elements serves different goals depending on the targeted audience. In addition, our study of potential communication of elements on the ethics of AI between mesostructures sets national and intergovernmental institutions as a high stake normative arena, where industrial and civil society actors fight to push their elements of language and impose their vision.

[21] Such as the GAFAMI – Google, Amazon, Facebook (Meta), Apple, Microsoft and IBM.

2.8.2 *Clusters draw a mesostructure of AI ethics*

Our clusters correspond to types of discourses around AI ethics that map onto existing mesostructures[22]. We acknowledge that, since we did not conduct a sectoral analysis, there is no evidence of which discourse directly map to which structure. However, we reckon that our clusters clearly draw themes that are usually associated with types of actors, such as the business cluster for companies or the regulatory clusters for institutional actors. We subsequently analyse the possible exchange of information between these mesostructures through AI ethics documents.

We examine three main types of actors: the AI industry, inter- and national governmental institutions and the civil society. These groups provide a grid for the conflicts and collaborations that necessarily arise when defining the outline of AI ethics. It is complementary to the recent work of Benbouzid et al. (2022) that outlined four normative arenas in AI: the transhumanist arena, the researchers trying to "solve" AI through technical improvements, the social critique of AI, and the regulation of AI markets. Here, we analyse how different structures may interact with each other, offering a complementary perspective to their work.

Let us emphasise that we focus on the mesostructures, and not the individuals: the same individual can (and does) take multiple roles depending on the context: an AI engineer can also be involved in civil society activism, for example, bringing their knowledge and skills to this sector, or sit on committees within regulatory institutions. These interactions are the mark of different commitments and multilateral dependencies, both moral and symbolic. Actors use formal organisations as tools, through conflicts and agreements, in order to set up and maintain social processes of solidarity, control, institutionalisation, etc. Those processes exist *because* of the existence of these organisations. Understanding these interactions is key in comprehending the social mechanisms that permit collective action, and by extension the dynamic landscape of AI ethics.

The technical and business clusters: the influence of AI companies

The industry communicates concrete elements, related to their technical models and applications, to national and international regulatory institutions, to nudge regulation in a way that benefits their business incentives. In particular, the AI industry is one of the main producers, with academia, of machine learning and artificial intelligence models, and build their business practices upon these[23]. Even in research, the state of the art in AI and the main discoveries often come from projects funded by the BigTech industry[24] As such, they have a vested interest in seeing forth to the development of said systems, and that shapes their interactions accordingly. This leads them to communicating abstract elements to conceptually explain their vision and strategy, which end up in our business cluster of terms. Their is no doubt that the communication of the more abstract elements about business strategies and of concrete elements about models and applications are first and foremost targeted at regulators, although business elements can also serve as a showcase for shareholders.

Governmental and European institutions

The role of public institutions, both national and international, is to provide normative directions, through laws and legal norms, to define the setting in which AI

[22] Mesostructure refers to *"the social processes and ordering that occur between the macro and micro levels of social organisation"* (Ulmer, 2015).

[23] For example, Microsoft invested 1 billion US dollars in OpenAI in 2019.

[24] Such as the transformer architecture (Vaswani et al., 2017) from Google.

is deemed acceptable. Regulatory bodies consist of an interesting arena, forming a middle ground between industry and the civil society, and subject to pressure from both sides. They take as input both abstract and concrete elements communicated by these entities, and produce new abstract elements, relating to a global vision, or concrete elements, as a stepping stone towards international and national regulation. This is most notably shown by the two regulatory clusters which highlight already existing law and rights, such as the GDPR and personal data protection, as well as means to implement regulation, such as with impact assessments, audit and accountability. We therefore expect these two regulatory clusters to be reinforced when looking at documents from these institutions specifically. But, on the contrary, we note that this is almost the other way around as a diluted legal cluster faces three clusters – business, technical and biometrics – comprising of elements that probably emanate from industrial stakeholders. While legal protection remains a central theme in European documents, it is strongly intertwined with practical considerations, demonstrating the influence of external incentives.

The civil society fails to address social concerns

The civil society comprises of every individual or group of individuals that is not from the private sector or public institutions. AI ethics documents emanating from this sector could therefore reveal various realities.

Nonprofits, for instance, are large structures which can put pressure on state and regulatory representatives. We noticed this pressure-inducing role in our previous study (Gornet and Viard, 2023a): after giving technical keys for their readers, documents emanating from the civil society tend to acutely describe the limitations of AI models and their deployment, and are less prudish about discussing power struggles along class, gender and race lines. In that sense, their output is tilted towards more concrete elements, specifically related to the failings of AI systems, while contributing abstract elements in the form of societal directions. On the contrary, documents written by small groups of individuals, who have little pressure means on the other sectors, are either mere echo chambers of the AI industry's tenets, or focus of a social critique of AI. Finally, academic structures employ scholars from both technical and social fields, that have deep and broad knowledge of the issues raised by artificial intelligence. As such, academic documents typically communicate abstract elements – for example about technical intricacies or social challenges, describing AI as a socio-technical system. They therefore provide abstract explanations and reasoning rather than concrete guidelines for decisions and can serve as a counterbalance to industry claims, having a less vested interest in selling AI models.

However, these elements on the social critique of AI, that are nevertheless present in charters[25], are here almost absent of our dataset. There can be several reasons to this. Firstly, we included in our dataset only documents that are cited in previous meta-analyses, therefore excluding some alternative frameworks[26]. Secondly, even when these documents appear in our dataset, their vocabulary is more specific as they deal with themes that only they address[27]. As a result, they are diluted by the statistical processing of documents and do not show up clearly in our clusters. The general discourse around AI ethics therefore tends to converge outside the sphere of social critique.

[25] From our own readings and from the results of our previous study (Gornet and Viard, 2023a).

[26] Such as the "Decolonial AI Manyfesto" (Krishnan et al., n.d.).

[27] For instance, "sexism" or "structural racism".

The necessity to reconcile social and technical issues

The broad principles, identified in other meta-analyses, are far from being operational. Yet, focusing solely on technical tools to solve AI problems, as industry players do, risks leading to techno-solutionism. This leaves a gap between the two that is not concretely addressed in the corpus we studied.

According to Munn (2022), there is a need to go beyond ethical principles and *"think broadly about system of oppression and narrowly about auditing"*. This tendency to segment is already what we observe in the texts: some are focusing more on broad issues of AI and others solely on tools to solve those issues. Yet, even documents discussing broad issues do not go as far as addressing systemic injustices. Similarly, documents that discuss existing tools to mitigate technical issues, such as biases, assume that their use will solve social problems, like discrimination. There is a need to link those social and technical stances by building frameworks that consider the ecosystem and overall context of AI systems, to enable actions at different levels, beyond vague goals and applied computing. It will become necessary to determine how to evaluate and audit AI systems, how to measure their impact on work conditions, climate change or political polarisation, and decide whether a given technology is deemed safe to deploy or not, and under which conditions.

2.8.3 AI ethics charters as an introduction to AI regulation

AI ethics charters are often seen as the precursor of other forms of regulation. They came before the current legal frameworks and sometimes largely influenced their development, such as in the case of the AI Act in Europe[28], the world's first mandatory framework for regulating AI (European Parliament and Council, 2024b). Ethics principles, such as *fairness*, *transparency* or *human oversight* for instance, are clearly found in the AI Act, in Articles 10, 13 and 14[29]. Indeed, since the HLEG ethics guidelines in 2019, and then with the AI Act in 2021, the European Union has adopted an approach based on "trustworthiness"[30], which tries to reconcile the legal, social and technical stances. If Europe wants to regulate AI effectively, it will have to learn to reconcile these different worlds that usually talk past each other.

However, the dynamics that we have identified in ethics charters, with the European institutions at the centre of the games of influence, are also strongly present in subsequent forms of regulation, such as the AI Act (Wachter, 2024) and technical standards[31]. Europe must therefore be cautious when considering the discourse of industrial stakeholders, which dilutes moral principles and the legal protection of individual rights in a business- and technical-centred approach.

2.9 Conclusion

In this chapter, we collected and created the first public corpus of AI ethics related documents with their contents, rather than a list of documents matched to a reading grid. We showed that our corpus covers significant portions of most well-known previous studies, and we use it to confirm past results. In addition to a pre-trained model, it can be used to measure and quantify word embedding bias in such documents, using current debiasing methods (Gonen and Goldberg, 2019; Schröder et al., 2021). After shortly describing the corpus and the term frequencies, we quantitatively analysed it along two axes: we use textual analysis to highlight

[28] See Chapter 3.

[29] Note, however, that *fairness* is reduced in the AI Act to only one of its various dimensions: *data quality* and the reduction of biases. For more information, see Chapter 6, Section 6.6.

[30] To learn more about the "trustworthiness" approach to AI regulation, see Chapter 3, Section 3.2.2.

[31] See Chapters 7 and 8.

the main areas being discussed, and semantic graph analysis to identify points of controversy. We analyse both the main corpus and four subcorpus, as well as compare our results to previous works.

Let us now detail a few perspectives this work opens. The most straightforward one is linked to the corpus: adding new documents to the corpus is made easy, and since all our code is available, makes reproducing our work with more data accessible. Another interesting perspective would entail setting up a data visualisation platform, to search, visualise and explore the corpus's documents, making our corpus a valuable tool for a wider audience. An interesting perspective is to study the temporality of these documents and concepts, in particular to outline arbitrations that durably shaped AI ethics. We would also like to explore the polysemy of words used in the AI field, by applying more advanced natural language processing methods to analyse the corpus' semantic contents. Using the Abstract Meaning Representation (AMR) framework (Banarescu et al., 2013), we can extract semantic graphs from each of the documents in the corpus, and then apply methods from graph studies to the obtained semantic graphs in order to identify the underlying structures.

In addition, we provide leads for a potential sociological analysis of the discourses around AI ethics. If our mesoanalysis will need to be validated by a sectoral study of the documents, it already draws a possible map of structures where the power dynamics around AI ethics force different actors to communicate on abstract and concrete elements, in technical, business-oriented and regulatory discourses. Given the influence of AI ethics documents on subsequent mandatory law such as the AI Act, deciphering their main trends will help us understand the intent behind the legal requirements, as well as reflect on how these elements were passed on, and what was left aside in the process.

Chapter 3
The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation

Where does this chapter come from?

This chapter is an original contribution, written specifically for this thesis. It aims to provide background on the AI Act, what it contains in terms of requirements and implementation mechanisms, as well as its inspirations from ethical and policy documents and other legal texts.

It was released online as a standalone work:

> Gornet, M. (2024a). The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation. `https://hal.science/hal-04785519`

Chapter's abstract

What with the dangers of artificial intelligence for individuals and society, and the rapid evolution of these technologies, Europe has decided to take the lead by imposing strict requirements for the placing on the market of "AI systems". This new European law, adopted in June 2024, is better know as "the AI Act". The AI Act is based on a hierarchy of risks, where riskier systems will be subject to stricter obligations. While the AI Act is not the first law in Europe to be based on risk – the General Data Protection Regulation (GDPR) and subsequent laws on digital technologies have already started this trend – it is the first to take it to such a level. But the AI Act also draws on the concept of "trustworthy AI", a term coined by policy documents that preceded it, and according to which AI must notably be ethical and technically robust.

In this chapter, we retrace the story of the AI Act, in order to understand the origin of its main concepts and structure. We also take a look at the final version of the text, its hierarchy of AI systems and the corresponding obligations, as well as the governance ecosystem it puts in place to ensure that these rules are properly implemented. The picture we draw shows a regulation that is quite unique in the European legal landscape, despite its many roots and inspirations.

3.1 Introduction

The European AI Act is the first mandatory framework adopted for AI in the world. At the time of writing, the final text has been published in the Official Journal of the European Union (OJEU) (European Parliament and Council, 2024b) and entered into force on 2 August 2024, although some requirements will apply later. This official endorsement follows months of negotiations between the three European institutions, the Commission, the Council and the Parliament, during which the latter two each suggested amendments to the text initially proposed. The European Union prides itself in this achievement, which took years in the making. Indeed, while the first proposal of the AI Act was published in 2021, the idea of developing a mandatory framework for AI in Europe is much older, being mentioned in policy documents dating back to 2017. What makes the AI Act's approach so different from other European legislation?

We start in Section 3.2 by recounting the story of the AI Act, from the first discussions on creation of a legal status for robots, through the recurring discourse on AI "trustworthiness" and "risk", to the final process of adoption of the text we know today. In Section 3.3, we show that the AI Act is not alone in the legal landscape of digital technologies in Europe and builds on previous frameworks for data protection, digital platforms, product safety, product liability and so on. In Section 3.4, we analyse one of the legal texts that strongly inspired the AI Act: the General Data Protection Regulation (GDPR). We show that the GDPR laid the foundations for a risk-based regulation in which technical standards would play a prominent role, although it did not go as far as the AI Act in this respect. After examining its various sources of inspiration, we take a look at the AI Act's provisions in Section 3.5: its scope, the classification of AI systems and their relative obligations, as well as its proposal for ensuring innovation and monitoring the proper enforcement of requirements. We also map the new governance ecosystem created by the AI Act, as well as the expected deliverables, and indicate important future dates for its entry into force. Finally, in Section 3.6, we look at the various criticisms that were addressed by the academic literature to the AI Act. While disagreements on the definition of AI, which systems should be prohibited or high-risk and how to implement requirements, are not entirely solved, they are not structural problems. On the contrary, we discuss in Section 3.7 the risk-based approach chosen by the Commission and how it also attempts to take fundamental rights into account. We show that the distinctive European approach to AI regulation, which blends risks and rights, raises questions about the implementation of the regulation.

3.2 The growing discussion on AI in Europe

In this section, we present a brief history of the policy documents and ethical guidelines published by the European institutions prior to the AI Act and how they influenced it. To guide the discussion, a timeline is presented in Figure 3.1.

3.2.1 AI increasingly became a topic of interest

In 2017, discussions on robotics reached the European institutions with the EU Parliament resolution on Civil Law Rules on robotics (European Parliament, 2017). The text was subsequently strongly criticised by AI and robotics experts, who were

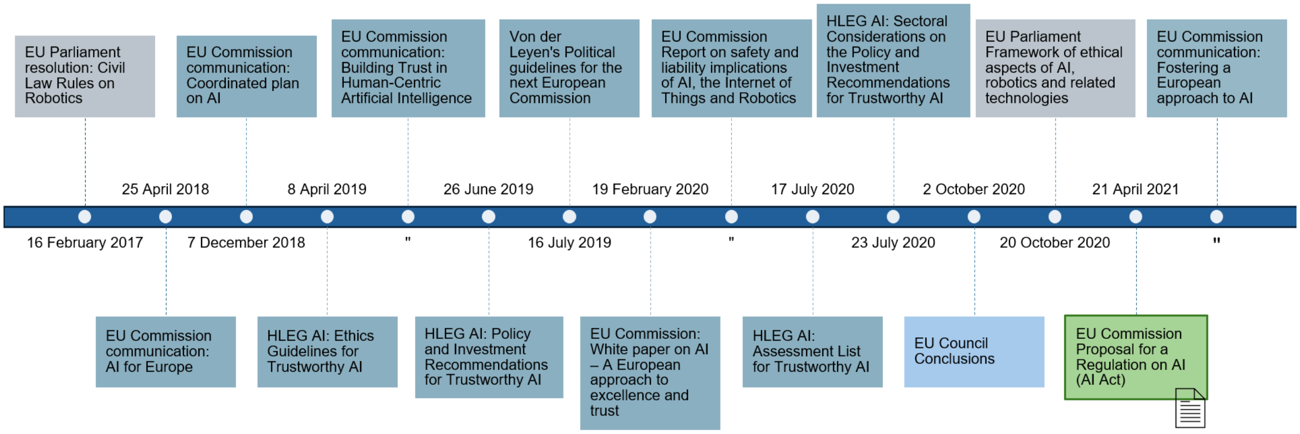

EU Parliament resolution: Civil Law Rules on Robotics
EU Commission communication: Coordinated plan on AI
EU Commission communication: Building Trust in Human-Centric Artificial Intelligence
Von der Leyen's Political guidelines for the next European Commission
EU Commission Report on safety and liability implications of AI, the Internet of Things and Robotics
HLEG AI: Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy AI
EU Parliament Framework of ethical aspects of AI, robotics and related technologies
EU Commission communication: Fostering a European approach to AI
25 April 2018
8 April 2019
26 June 2019
19 February 2020
17 July 2020
2 October 2020
21 April 2021
16 February 2017
7 December 2018
"
16 July 2019
"
23 July 2020
20 October 2020
"
EU Commission communication: AI for Europe
HLEG AI: Ethics Guidelines for Trustworthy AI
HLEG AI: Policy and Investment Recommendations for Trustworthy AI
EU Commission: White paper on AI – A European approach to excellence and trust
HLEG AI: Assessment List for Trustworthy AI
EU Council Conclusions
EU Commission Proposal for a Regulation on AI (AI Act)

Figure 3.1: Timeline of policy documents and ethical guidelines published by the European institutions from 2017 to 2021. Representation from the author.

particularly concerned about one of the European Parliament's recommendations to the EU Commission, pushing for the creation of a legal status for robots in the long run. For the experts, giving robots legal status was a slippery slope as it would have grant them rights and obligations, blurring the lines between science fiction and reality, and opening the door to liability issues, as any accident caused by the robot would have incurred the liability not of its owner, but of the robot itself (Robotics Openletter, 2017). This idea of giving legal status to a robot or algorithm was latter abandoned and publications started focusing more and more on AI. Since then, AI and digital technologies have been at the top of Europe's agenda. Ursula von der Leyen, then candidate for the presidency of the European Commission, made it one of her priorities to have *"a Europe fit for the digital age"* (von der Leyen, 2019).

But while the EU Commission has revealed a plethora of different AI policy documents since then, the EU Council only mentions AI in the conclusions of a meeting held in October 2020, where it recognises that the EU needed to be *"a global leader in the development of secure, trustworthy and ethical AI"* (European Council, 2020). It further called on the EU Commission to propose ways of developing research and innovation in the field of AI. The European Parliament, on the other hand, has steered its course on AI through a series of resolutions, generally focusing on sector-specific measures, such as criminal justice or education, and on specific issues raised by AI, such as intellectual property rights or the economic aspects of AI (European Parliament, n.d.a).

In April 2018, the European Commission presented its strategy for AI in a communication entitled "Artificial Intelligence for Europe" (European Commission, 2018a). This strategy was intended as a response to the rapid progress being made by the United States and China in the field of AI, with both countries battling for the lead in the "AI race", with Europe clearly lagging behind (Smuha and Yeung, 2024). The European approach to AI regulation was shaped as a distinct brand, based on European values, positioning itself in contrast to the state-controlled model of China and the permissive model of the US.

The strategy presented relied on four key points moving forward: (1) *"boosting the EU's technological and industrial capacity"* by increasing investments in AI, supporting research, building testing infrastructures and making more data available; (2) *"preparing for socioeconomic changes"* by encouraging diversity and interdisciplinarity and creating an attractive environment for talents in the EU; (3) *"ensuring an appropriate ethical and legal framework"* by drafting AI ethics guidelines and ensuring safety and liability; and (4) *"joining forces"* by engaging both with Member States and AI stakeholders. This communication is at the origin of a number of initiatives that we know today. In particular, it encouraged the drafting of AI ethics guidelines, which would later lead to the creation of the High Level Expert Group on AI (HLEG). The communication also stressed the need to develop standards to *"increase consumer trust"*. As a result, standards now play a major role in the AI Act[1]. Finally, the communication discusses the need to reinterpret the Product Liability Directive in light of technological developments, leading to a proposal for a revision of the framework in 2022 that will later be adopted (European Parliament and Council, 2024a), as well as a proposal for a new AI Liability Directive, still under discussion (European Parliament and Council, 2022).

3.2.2 *A European discourse based on trust and respect for fundamental rights*

A second communication from the Commission was published in December of the same year, the "Coordinated plan on AI" (European Commission, 2018b), containing actions to be undertaken by the Member States and the Commission. Investment and support resources were specified, with quantified objectives. In particular, a deadline was set for the development of ethical guidelines in March 2019. Both communications emphasised the need for legal rules and ethics guidelines, meant to complement each other and to help protect fundamental rights. The emphasis is on put on "trust", which will later become the cornerstone of all EU deliverables on AI.

Before the legal rules of the AI Act, the ethical framework was the first to be put in place, with the creation of the European Commission's High Level Expert Group on AI (HLEG) in June 2018. The HLEG was tasked with preparing two complementary deliverables: one aimed at AI practitioners, the "Ethics Guidelines for Trustworthy AI" (HLEG, 2019a), and the other addressed to the EU institutions and Member States, the "Policy and Investment Recommendations for Trustworthy AI" (HLEG, 2019b). On the day of release of the Ethics Guidelines, the EU Commission published its third communication on AI: "Building Trust in Human-Centric Artificial Intelligence" (European Commission, 2019). The aim of this last communication was to support the work of the HLEG, by summarising the experts' conclusions and outlining the next steps in updating these guidelines and exporting the EU's expertise in drafting ethical guidelines outside Europe.

Subsequently, the HLEG Ethics Guidelines became one of the most referenced AI ethics documents [2]. In many ways, these guidelines have greatly influenced the field of AI ethics, as well as the discourse of European institutions going forward. The guidelines can therefore be seen as the cornerstone of the European strategy for regulating AI. Notably, they introduced the term "trustworthy AI", which will remain in all of the following European documents, including the AI Act. According to the guidelines, for AI to be considered "trustworthy", it should be: lawful, ethical, and robust.

[1] See Chapter 4, Section 4.7.

[2] Based on the study conducted for Chapter 2.

> ***Definition***
>
> ***Trustworthy AI (HLEG 2019).*** *Trustworthy AI has three components, which should be met throughout the system's entire life cycle: 1. it should be lawful, complying with all applicable laws and regulations; 2. it should be ethical, ensuring adherence to ethical principles and values; and 3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.*

This enables us to identify the three pillars that the EU has chosen to push forward: the legal sphere with the AI Act, the ethical sphere with ethics guidelines, and the technical sphere with standards and product safety. These pillars, as intended by the Commission, are represented in Figure 3.2.

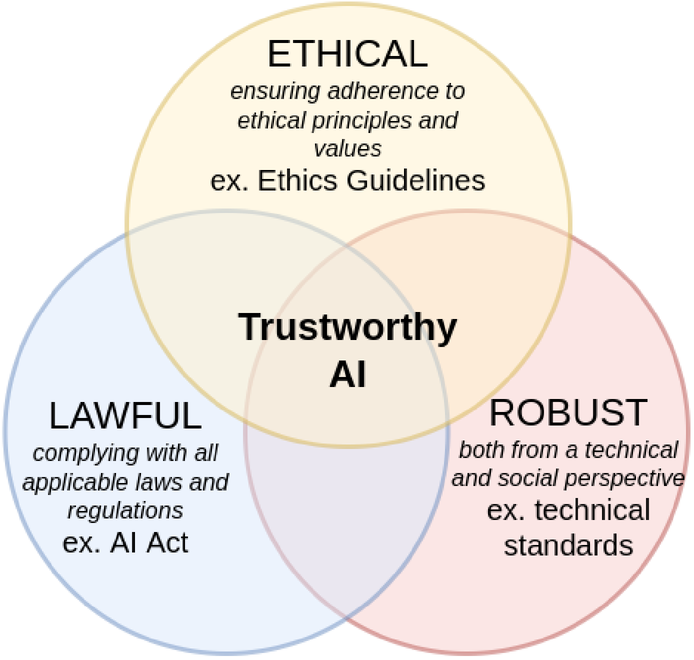

ETHICAL
ensuring adherence to ethical principles and values
ex. Ethics Guidelines
Trustworthy AI
LAWFUL
complying with all applicable laws and regulations
ex. AI Act
ROBUST
both from a technical and social perspective
ex. technical standards

Figure 3.2: Visual representation of the three pillars of "Trustworthy AI", as intended by the HLEG. The three spheres – lawful, ethical and robust – are studied in the thesis through respectively ethics guidelines, the AI Act and technical standards.

After a year's absence, the HLEG was back in force in 2020 with two new publications: the "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment" (HLEG, 2020a), and "Sectoral Considerations on the Policy and Investment Recommendations" (HLEG, 2020b). Both of these documents were follow ups of their previous documents – the Ethics Guidelines and the Policy and Investment Recommendations. The Sectoral Considerations were primarily geared towards the industry providing AI in three sectors – the IoT sector, the public sector and healthcare, while the ALTAI attempted to overcome one of the biggest challenges of ethical guidelines: operabilty. Indeed, aware that the other charters were merely lists of inoperative principles, one of the HLEG's objectives was to go beyond simply listing ethical principles and to provide guidance on their practical implementation (Smuha, 2019). Alongside the three components for "trustworthy AI" – lawful, ethical and robust, the Ethics Guidelines therefore introduced an additional list of seven key requirements: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability. Each of these requirements was then dissected in the ALTAI into a series of questions addressed to AI practitioners in companies. The ALTAI is designed as a checklist, a tool to support the development of "trustworthy AI" For each requirement, questions are asked on the context in which the system will be deployed, as well as the processes or measures put in place to take the requirement into account. The precision and technicality of the questions asked make the ALTAI a special AI ethics document, halfway between ethical guidelines and technical standards.

The Ethics Guidelines and the ALTAI have notably influenced some of the requirements of the AI Act, such as one of the amendments proposed by the EU Parliament which introduced a Fundamental Rights Impact Assessment (FRIA). At the time the amendment was proposed, the HLEG's "key requirements" were included directly, without change, in the list of criteria for assessing fundamental rights[3].

3.2.3 *The birth of a risk-based approach*

The second deliverable of the HLEG, the Policy Recommendations, also had a major influence on the drafting of the AI Act. Indeed, for the first time, the document called EU institutions to adopt *"a risk-based approach to policy making"*, but insisting also on a complementary *"precautionary principle-based approach"* that was later tuned down by the Commission in the proposal for an AI Act (Smuha and Yeung, 2024). Notably, although the HLEG proposed the risk-based approach, it did not advocate the use of the product safety framework for AI (Almada and Petit, 2023). This approach came directly from the Commission when it proposed the AI Act[4].

The Commission continued to emphasise the need for more regulation through two subsequent publications: the "White Paper on AI" (European Commission, 2020d) and a report on the "safety and liability implications of AI" (European Commission, 2020c), both published in February 2020. The report, while acknowledging that safety and liability frameworks already applied to AI products, noted that some risks specific to these technologies were not addressed by current frameworks and that *"additional obligations may be needed"*. The White Paper claimed that the product safety and civil liability legislation was insufficient at the time to deal with AI-related damage, and called for these frameworks to be supplemented. Furthermore, the risk-based approach introduced in the Policy Recommendations and the emphasis on trust from the HLEG guidelines were enhanced by the Commission in the White paper.

The White paper proposed two categories of risks: *"high risk"* and *"not high risk"*. For high-risk AI systems, the White paper lists some provisions that would subsequently become requirements under the AI Act, such as the obligation to use representative datasets, provisions on record keeping, transparency, robustness and accuracy, human oversight, etc. In the White paper, the Commission also recognises the need for prior conformity assessment, which would verify and ensure that the requirements for high-risk applications are complied with. The White paper notably states that: *"the prior conformity assessment could include procedures for testing, inspection or certification. It could include checks of the algorithms and of the data sets used in the development phase"*; and that *"the conformity assessments for high-risk AI applications should be part of the conformity assessment mechanisms that already exist for a large number of products being placed on the EU's internal market"*. Furthermore, standards are cited in the White paper as a means of facilitating compliance. It is therefore

[3] This provision is included in (European Parliament, 2023a, amendment 413). In the final text of the AI Act, the list of criteria is not included, but the obligation to carry out a FRIA for high-risk systems is maintained (Art. 27 AI Act).

[4] See Section 3.2.4 of this chapter.

clear that the risk-based structure of the AI Act, as well as parts of its provisions on conformity assessments and standards, are inherited from the White paper[5].

In addition, for AI applications that would not qualify as high-risk, the White paper proposes the use of voluntary labels. Although this is not the solution chosen in the AI Act, it comes close to the use of codes of conducts for "other AI systems"[6]. Finally, the White Paper also stresses the need for a European governance framework, based on national authorities, but also on participation and advice from various stakeholders. The AI Act took these ideas and improved on them, including the creation of new entities which are presented in Section 3.5.4.

Following the Commission's documents, the European Parliament strengthened the case for more regulation in October 2020 by adopting two Resolutions. In its first Resolution on a "civil liability regime for AI", the Parliament recognised that while a *"complete revision"* of the liability regime was not needed, *"adjustments"* were necessary (European Parliament, 2020b). The Resolution includes a Proposal for a Regulation that has served as inspiration for the Commission's proposal for an AI Liability Directive. In the second Resolution on a "framework of ethical aspects of artificial intelligence, robotics and related technologies", as in the Resolution on a liability regime for AI, the Parliament included a draft proposal for a Regulation (European Parliament, 2020c). Although the framework has been considerably modified by the Commission, the Parliament's proposal is at the origin of what will become the AI Act. However, unlike the Commission's AI Act, and as the title of the Resolution suggests, the Parliament's proposal is strongly based on ethical principles and fundamental rights, such as human oversight, transparency and non-discrimination. Yet, the Parliament seems to mix ethical principles and fundamental rights without making a clear distinction: Article 5 of the proposal is entitled "ethical principles of AI" but deals directly with the fundamental rights enshrined in the EU Charter. In addition to an approach based on ethics and fundamental rights, the Parliament also advocates in the Resolution for a risk-based approach to AI regulation, where compliance would be based on standards. The Resolution states that *"any future regulation should follow a differentiated and future oriented risk-based approach to regulating artificial intelligence, robotics and related technologies, including technology-neutral standards across all sectors, with sector-specific standards where appropriate"*. The Parliament, like the Commission in the White Paper, refers to high-risk AI applications that would be subject to mandatory compliance.

3.2.4 *Towards a mandatory horizontal regulation*

The various policy documents published by the three EU institutions – the Commission, the Parliament and the Council – have increasingly moved away from a discourse based on ethics and fundamental rights towards strict regulation based on the risks posed by AI, compliance with which could be assessed by means of technical standards. It was this approach that gave rise to the AI Act. But to explain this transition from ethics, trust and fundamental rights to what is now a product safety regulation, the Commission has released a number of documents alongside the AI Act proposal in April 2021.

First, like all EU regulation, the AI Act was accompanied by an "impact assessment" (European Commission, 2021b), designed to explain the Commission's choice to adopt a certain regulatory approach and why other possible approaches were rejected. Five options were initially considered: (1) a EU voluntary labelling

[5] For more information on the risk-based structure of the AI Act, see Section 3.5.2. For more information on conformity assessment and standards under the AI Act, see Chapter 4, Section 4.7.2 and Section 4.7.3.

[6] See Section 3.5.2.

scheme, (2) an ad-hoc sectoral approach, (3) an horizontal risk-based act, (4) codes of conducts, and (5) an horizontal act for all AI. The impact assessment describes each option and further looks into the advantages and disadvantages of each of them, based on certain themes. In almost all themes, options 3 and 4 are favoured, and the others are found to have too many drawbacks. The labelling scheme option is considered uncertain, with no guarantee that it will be widely adopted, as companies will only agree to undergo a labelling audit if the costs are lower than the benefits. The sectoral approach was also rejected on the grounds that it could lead to inconsistencies in the requirements imposed by sectoral legislation, that regulation would only take place once concerns had been identified, that it would not prevent Member States from adopting their own horizontal regulations, leading to heterogeneous legislation across the EU, and that companies with different AI systems used for different use cases would have to bear multiple compliance costs. As for the last option, which would impose the same strict requirements on all AI systems, although the Commission recognises that it would protect thoroughly the safety and fundamental rights of citizens, it would also expose small businesses to potentially significant compliance costs and create a heavy regulatory burden. On the contrary, the option of a horizontal regulation based on different categories of risk, completed by codes of conduct for low-risk systems, is supposed to *"enhance users' trust"*, and increase legal certainty. However, the core of the impact assessment focuses on the risk-based option, which is the most widely discussed[7]. It therefore seems like the other less developed options were never seriously considered.

In addition to the impact assessment, the proposal on an AI Act also came with an introductory part: the Explanatory memorandum, which offers context, reasons for the proposal and allow us to understand the intent of the Commission[8]. Several of the documents mentioned earlier in this section – such as the HLEG deliverables, the White Paper on AI or the Parliament resolutions – are cited in the explanatory memorandum, as they were major sources of inspiration for the legal text. The Explanatory memorandum gives further arguments to justify the choice of an horizontal approach. We learn that the regulation is intended to be *"comprehensive and future-proof"* with *"flexible mechanisms that enable it to be dynamically adapted as the technology evolves and new concerning situations emerge"*.

These documents justify the European Commission's approach with the AI Act, which is not sector-specific, but risk-specific, with the desire to cover AI in general, but with different requirements depending on the risk category[9]. EU officials have declared that this horizontality guarantees *"functional equivalence"*, whereby obligations and enforcement tools are the same regardless of the sector in which the technology is used (Mazzini and Scalzo, 2023).

3.2.5 *The process of adoption of the AI Act*

On 21 April 2021, the European Commission published the first proposal to regulate artificial intelligence in Europe (European Commission, 2021e) which will come to be known as the AI Act. Figure 3.3 traces the timeline of the text's adoption, from this first proposal to the recent publication of the AI Act in the OJEU.

Along with the AI Act, the Commission launched a consultation period, where stakeholders from various backgrounds[10] were given the opportunity to provide comments on the text. The Commission received over three hundred submissions[11].

The release of the first draft of the text by the EU Commission was the first step

[7] The description of all the options runs from page 39 to page 62. The risk-based approach is described on 14 pages, while the other options are described much more briefly, the maximum being the sectoral option, which is described on 5 pages.

[8] The Explanatory memorandum appears on the same document as the Commission proposal and can thus be found here (European Commission, 2021e).

[9] For more information on these different categories, see Section 3.5.2.

[10] Notably NGOs, academic and research institutions, companies and businesses, and various civil society actors.

[11] All submissions for the AI Act can be found on the Commission's website: (European Commission, 2021a).

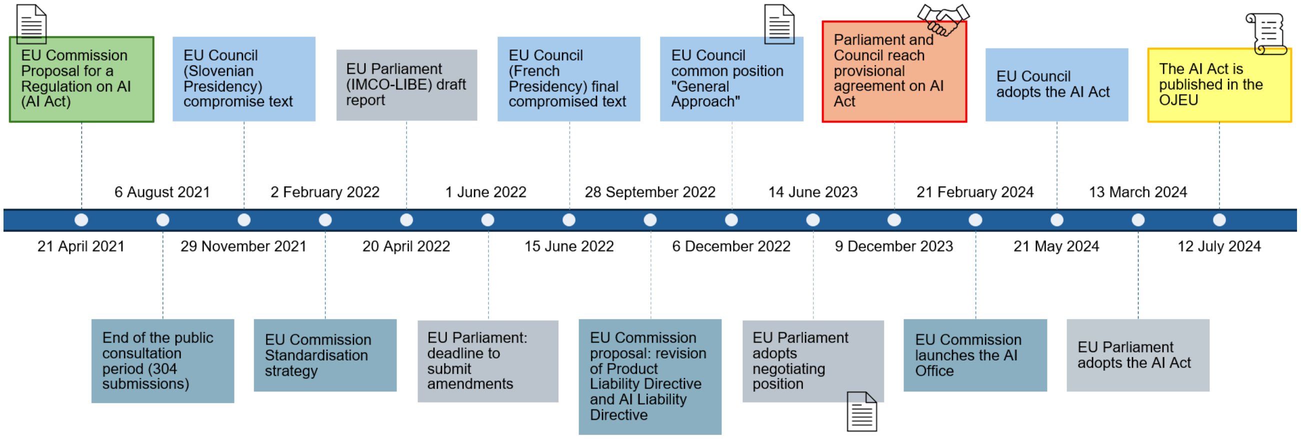

EU Commission Proposal for a Regulation on AI (AI Act)
EU Council (Slovenian Presidency) compromise text
EU Parliament (IMCO-LIBE) draft report
EU Council (French Presidency) final compromised text
EU Council common position "General Approach"
Parliament and Council reach provisional agreement on AI Act
EU Council adopts the AI Act
The AI Act is published in the OJEU
6 August 2021
2 February 2022
1 June 2022
28 September 2022
14 June 2023
21 February 2024
13 March 2024
21 April 2021
29 November 2021
20 April 2022
15 June 2022
6 December 2022
9 December 2023
21 May 2024
12 July 2024
End of the public consultation period (304 submissions)
EU Commission Standardisation strategy
EU Parliament: deadline to submit amendments
EU Commission proposal: revision of Product Liability Directive and AI Liability Directive
EU Parliament adopts negotiating position
EU Commission launches the AI Office
EU Parliament adopts the AI Act

Figure 3.3: Timeline of the adoption of the AI Act.

towards the adoption of a mandatory framework for AI in the EU. However, to come into force, the AI Act had still a long way to go. The two other EU institutions, the Council and the Parliament, had to propose amendments to the text. The rotating presidency of the Council meant that one member state would lead efforts to amend the text for six months before another member state took over. The first amendments were therefore proposed at the initiative of the Slovenian presidency and covered only Articles 1 to 7, making changes to prohibited and high-risk AI systems in particular (European Council, 2021). The next presidency, led by France, then proposed a large number of changes, notably to Article 4, proposing to regulate general purpose AI systems (European Council, 2022b). These contributions were brought together in the French presidency's compromise text (European Council, 2022c). The subsequent Czech presidency continued the process of amending the text, resulting in the General Approach, at the end of 2022 (European Council, 2022a)[12].

The European Parliament operates differently to the Council. The Parliament is made up of different committees, responsible for examining legislative proposals and proposing amendments, which are then submitted in the form of reports to the Parliament who adopts them in plenary session. Two of these committees were chosen to lead the negotiations on the AI Act: the Committee on the Internal Market and Consumer Protection (IMCO) and the Committee on Civil Liberties, Justice and Home Affairs (LIBE) (Ada Lovelace Institute, 2021). Five additional committees[13] adopted their own opinions, with proposed amendments to the AI Act. After an agreement was found between the different committees, the negotiating position was proposed to the Parliament as a whole and adopted on 14 June 2023 (European Parliament, 2023a).

After the adoption of the negotiating positions and proposed amendments, the three EU institutions – the Commission, Council and Parliament, entered a "trilogue" phase to discuss these amendments. After a 3-day "marathon" talk, the EU Commission, Council and Parliament reached a provisional agreement on the text on 9 December 2023 (European Parliament, 2023b). The text then underwent a series of minor textual improvements and was approved by the Parliament on 13 March 2024 (European Parliament, 2024) and by the Council on 21 May 2024 (European Council, 2024). The text of the AI Act was subsequently signed on 13 June 2024 and published in the OJEU on 12 July (European Parliament and Council, 2024b).

[12] A more precise timeline is given by (Future of Life Institute, n.d.).

[13] Namely, the Committee on Legal Affairs (JURI), the Committee on Industry, Research and Energy (ITRE), the Committee on Culture and Education (CULT), the Committee on the Environment, Public Health and Food Safety (ENVI) and the Committee on Transport and Tourism (TRAN)

3.2.6 *A continuity of the trustworthiness discourse after the AI Act*

Communications and policy documents on AI did not stop with the publication of the AI Act. A wide range of documents continue to be published and EU institutions continue to launched projects. These initiatives explain the European approach to AI regulation, picking up the main key elements and providing new details on the way forward.

First, the Commission released alongside the AI Act a Communication on "Fostering a European approach to AI" to summarise the main elements present is the legal text (European Commission, 2021c). The AI Act is notably said to *"combine greater safety and fundamental rights protection while supporting innovation, enabling trust without preventing innovation"*.

Other documents provide guidance for EU AI policy. For example, the revision of the "Coordinated Plan on AI" sets out the next steps of the EU's strategy for AI (European Commission, 2021d). It is described by the Commission as the *"next step in creating EU global leadership in trustworthy AI"*. The coordinated plan sets several goals: to enable AI development and uptake, foster research excellence, promote the EU vision of *"AI for people"* and as a *"force for good in society"*, and strengthen leadership in key sectors, such as environment, robotics, health, public sector, law enforcement, mobility and agriculture. In particular, since the first version of the coordinated plan in 2018 (European Commission, 2018b), the Commission is committed to opening a small number of *"specialised large-scale reference sites"* across Europe, equipped with technology infrastructures and specific expertise: the AI Testing and Experimentation Facilities (TEFs) (European Commission, n.d.u). Since then, a few collaboration projects have been launched.

In addition, the Commission also published in January 2024 a Communication on *''boosting startups and innovation in trustworthy AI"* (European Commission, 2024d). It describes new initiatives to support AI startups and SMEs, including the launch of *"AI Factories"*, i.e. computing facilities, resources and services to attract AI *"talents"*. Other initiatives include a number of research and investment programmes.

The work to successfully implement the AI Act will also continue with the AI Office, a new executive organ of the European Commission created by the AI Act[14]. Initiatives led by the AI Office include the "AI Pact", a voluntary framework towards the industry to anticipate and prepare for future compliance with the AI Act (European Commission, n.d.b). A first call of interest was launched in November 2023. The AI Office then released the AI Pact commitments in September 2024, inviting participating companies to endorse this non-binding framework and report on their progress later. By signing up this Pact, the companies notably pledge to adopt an AI governing strategy, to identify their high-risk systems[15], and to promote AI literacy among staff.

3.3 *The legal landscape of the AI Act*

3.3.1 *Digital constitutionalism in Europe*

The AI Act will not apply in a vacuum. These past few years, the European Union has produced a proliferation of texts designed to regulate both new technological products and industrial players in the digital age. Some of these texts predate the AI Act, the best known being the GDPR (European Parliament and Council, 2016b)

[14] See Section 3.5.4.

[15] High-risk systems are a specific category of AI systems under the AI Act. For more information on the different categories, see Section 3.5.2.

for the protection of personal data and the Digital Services Act (DSA) (European Parliament and Council, 2022) and Digital Markets Act (DMA) (European Parliament and Council, 2022a) for the regulation of online platforms. Other texts are still in the making, such as the AI Liability Directive (European Parliament and Council, 2022), or the revision of the e-Privacy Directive (European Parliament and Council, 2017a). In total, there are dozens of texts which, if adopted, will regulate digital technologies in Europe, and the number of legislative proposals is likely to increase still further. [16] The objective of these texts is, among other things, to protect the fundamental rights of EU citizens, which is why scholars have been referring to this trend as *"digital constitutionalism"* (De Gregorio, 2021), i.e. an *"ideology that aims to establish [...] a normative framework for the protection of fundamental rights and the balancing of powers in the digital environment"* (Celeste, 2019). The AI Act is therefore part of this European approach to new technology regulation and will work alongside these other texts in the European legal landscape – some of which are quoted directly in the AI Act.

To navigate this legal landscape, we created a diagram showing the texts which are likely to intersect with the AI Act. This representation is illustrated in Appendix A.1. It shows all the treaties, directives and regulations cited in the AI Act, organised into large families corresponding to thematic spheres. The diagram is not intended to be an exhaustive list, but rather to give an idea of the multitude of texts involved. In the next section, we take a look at a selection of important texts.

3.3.2 *Relative treaties, directives and regulations*

The EU treaties

The AI Act, like all European legislation, is based on the EU Treaties. The two core treaties of the EU, the Treaty on European Union (TEU) (European Commission, 2012b) and the Treaty on the Functioning of the European Union (TFEU) (European Commission, 2012c), define how the EU operates. In accordance with the ordinary legislative procedure established by the Treaty of Lisbon, all new EU legislation must have a legal basis from one of the articles of these treaties[17]. For the AI Act, the Commission motivated the proposal on the basis of data protection (Art. 16 TFEU), and functioning of the internal market (Art. 114 TFEU). The Charter of Fundamental Rights of the European Union (CFREU) (European Commission, 2012a) is an additional text to be taken into consideration, especially as AI systems represent a danger for individuals and society as a whole.

Personal data protection

Europe also lead the way in personal data protection regulation with the GDPR (European Parliament and Council, 2016b) in 2016. The text lays down rights for data subjects and obligations from data controllers and data processors. Data must notably be processed in a transparent and secure manner, and for limited purposes[18]. In the context of law enforcement, the GDPR does not apply, but a second text, usually generally to as the "law enforcement directive" (European Parliament and Council, 2016a) takes over. Finally, a third text lays down obligations for data processing by the European institutions (European Parliament and Council, 2018).

[16] See a list of future European texts at (Zenner et al., 2024).

[17] See a list of these legal bases in (European Parliament, n.d.b).

[18] For more information on the GDPR and how it served as a source of inspiration for the AI Act, ses Section 3.4.

Data sharing

But not all data is personal and is covered by the GDPR. To facilitate the sharing in Europe of industrial data and notably data from the Internet of Things (IoT), the European Union adopted in 2022 and 2023 the Data Governance Act (European Parliament and Council, 2022c) and the Data Act (European Parliament and Council, 2023b). In particular, these laws allow data from connected devices to be made accessible, provide for the use of company data by public bodies in exceptional circumstances, abolish fees for changing service providers and offer a number of guarantees against illicit access to data by the governments of third countries.

Digital platforms

After the GDPR and data protection, the European institutions have been tackling the issue of the major digital platforms, with the publication of the DMA (European Parliament and Council, 2022) and DSA (European Parliament and Council, 2022a) in 2022. But the two texts serve very different objectives. The goal of the DMA is to complement competition law to prevent the monopoly of very large platforms – social networks, web browsers, etc. In particular, it aims to facilitate unsubscribing and interoperability with competitors, prohibits self-referencing, forces platforms to inform users on future fusions, and requires consent for the re-use of personal data for targeted advertising purposes. On the other hand, the DSA is focused on internet service providers, cloud services and online platforms. It aims to harmonise regulations on illegal content and products, such as hate speech, child pornography, terrorism, disinformation, drugs, counterfeit goods, and so on. It requires platforms to have a tool for reporting content and handling complaints. It provides for a right to explanation of algorithms, prohibits advertising targeted at minors and provides for risk analysis and annual audits and (limited) access to the interface.

Product safety

The AI Act is a product safety regulation, part of the New Legislative Framework (NLF) [19]. AI products are therefore *de facto* covered by the General Product Regulation and, for systems which do not fall into specific categories under the AI Act[20], they will at least be covered by the requirements of their sector-specific regulation. At the time of writing[21], there are 27 directives and regulations aligned with, or based on the NLF[22]. One key inspiration for the AI Act was the Medical Device Regulation (European Parliament and Council, 2017b), which is broadly seen by Europe as the success story of product safety regulation (Mazzini and Scalzo, 2023). Other example of product safety regulations include for instance the Toy Safety Directive (European Parliament and Council, 2009b), the Machinery Regulation (European Parliament and Council, 2023a) and the Radio Equipment Directive (European Parliament and Council, 2014).

Product liability

The AI Act is an *ex ante* regulation: it sets out the requirements that AI systems must meet *before* being placed on the market, but it does not cover *ex post* liability rules. Pending the specific rules set out in the AI Liability Directive (European Parliament and Council, 2022), the AI Act will for the time being follow the

[19] For more information on the NLF, see Chapter 4, Section 4.6.

[20] See Section 3.5.2.

[21] November 2024.

[22] A list of these directives and regulations can be found at (European Commission, n.d.r).

Product Liability Directive (European Council, 1985a). This Directive has recently been revised to integrate consideration for digital products, notably AI (European Parliament and Council, 2024a).

3.4 *The GDPR: the predecessor for digital rights*

3.4.1 *A risk-based approach initiated by the GDPR*

Among the texts of the European legal ecosystem, one stands out: the GDPR. As the first legal text adopted in Europe to tackle digital technologies directly, the GDPR has strongly influenced AI Act in its spirit and structure, and paved the way for future digital laws.

And indeed, the AI Act takes inspiration from the GDPR on a number of points. The GDPR's main objective is to protect citizens' rights to privacy and data protection. But while protecting these rights, the GDPR has also introduced the beginning of a risk-based approach that will be taken up and enhanced in the AI Act.

In the GDPR, the data controller[23] must carry out a Data Protection Impact Assessments (DPIAs) for high-risk processing operations (Art. 35(1) GDPR). This includes profiling, large scale processing and systematic monitoring (Art. 35(3) GDPR). The assessment must contain a description of the operations, an assessment of the necessity and proportionality of operations, an assessment of the risks to rights and freedoms of data subjects, and the measures envisaged to address these risks (Art. 35(7) GDPR). When the DPIA shows that the processing presents a high-risk in the absence of mitigating measures, the supervisory authority must be consulted (Art. 36(1) GDPR). The GDPR's DPIAs have notably influenced the FRIAs of the AI Act.

In addition to its novel risk-based approach, the GDPR has initiated a shift from a system of static prior formalities to a system of dynamic global compliance. Businesses do not need prior authorisation from supervising authorities to process personal data as before, but instead must be able to demonstrate at any time that they are complying with the principles of the Regulation. This is best shown in Article 24 on the responsability of the controller where they must *"ensure and be able to demonstrate that processing is performed in accordance with [the] Regulation"* by means of technical and organisational measures. Recital 74 goes even further, asserting that this obligation to demonstrate compliance also applies to the effectiveness of the measures. The notion of compliance in the GDPR is thus very much intertwined with the notion of risk[24], the lack of compliance creating more risks to the data subjects' right to privacy.

3.4.2 *The importance of the state of the art in the GDPR*

The state of the art plays an important role in the GDPR, a role that will be further strengthened by the AI Act by relying on harmonised standards[25]. Indeed, the GDPR makes trade conditional to the fulfillment of certain obligations by the controller, which are directly defined in the text. For instance, under the principle of integrity and confidentiality, the controller must ensure *"appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage"* (Art. 5(1)(f) GDPR). However, the means of achieving this goal in practice are left to the controller, who must take *"appropriate*

[23] The controller is the natural or legal person who determines the purposes and means of the processing of personal data (Art. 4(7) GDPR).

[24] Gellert (2018) speaks of a *"compliance risk"*.

[25] See Chapter 4, Section 4.7.

technical or organisational measures", but none are cited in the text. This security obligation is further strengthened by Article 32 on the security of processing, which stipulates that *"the controller and the processor*[26] *shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk"*. Unlike Article 5, however, Article 32 gives broad examples of how these measures can be carried out, such as the pseudonymisation and encryption of personal data. Yet, these measures remain very general and no concrete information is given on the technical aspects of their implementation. It is up to the controller and processor to choose what measures to put into place to ensure and demonstrate compliance with the Regulation (Art. 24(1) and 28(1) GDPR).

One key requirement introduced by the GDPR is that privacy and data protection need to be integrated in the technology when it is created, which is referred to as data protection "by design" and without human intervention, known as data protection "by default" (Art. 25 GDPR). This approach by design is also included in the AI Act, particularly in the requirements for high-risk systems: mitigating risk management requires *"adequate design and development"* (Art. 9(4) GDPR) and data management includes *"relevant design choices"* (Art. 10(2)(a) GDPR). Furthermore, high-risk systems must be *"designed and developed"* to enable the recording of events (Art. 12(1) GDPR), the transparency of the system and the interpretation of outputs (Art. 13(1) GDPR), an effective oversight (Art. 14(1) GDPR) and an appropriate level of accuracy, robustness and cybersecurity (Art. 15(1) GDPR). This approach by design notably requires consideration of the state of the art, as noted by Article 25(1). For instance, the processing security measures required by Article 32 must be taken into account the state of the art.

3.4.3 *Standards for the GDPR*

However, this does not mean that there are no technical standards for the GDPR. The most well known standard to tackle issues close to that of the GDPR is ISO/IEC 27701 (2019), based on two information security standards: ISO/IEC 27001 (2022a) on information security management systems and ISO/IEC 27002 (2022b) on security measures. However, while applying these standards is a way for companies to show that they are implementing good practices and can be seen as a first step towards compliance with GDPR requirements (Lopes et al., 2019), it remains insufficient to demonstrate full compliance with the European law. In particular, the French data protection authority, the CNIL, has stated that these standards are not GDPR specific and cannot be considered as a valid certification scheme for the GDPR, although they do represent the state of the art (CNIL, 2020). Indeed, ISO standards are international standards, far removed from European concerns about privacy protection. The first version of ISO/IEC 27001 for instance was published in 2005 (ISO/IEC, 2005), well before the GDPR was adopted in Europe.

To bridge the gap between insufficient international standards and GDPR requirements, the European Commission mandated the European standardisation organisations to prepare standards for the GDPR in a standardisation request on "Privacy and personal data protection management" delivered in 2015[27]. A few standards have been developed following that request[28]. These include standard EN 17529:2022 on "Data protection and privacy by design and by default" developed by CEN-CENELEC Joint Technical Committee 13 on "Cybersecurity and Data Protection". However, this standard, although adopted at European level, is not

[26] The processor is the natural or legal person who processes personal data on behalf of the controller (Art. 4(8) GDPR).

[27] The reference to such request can be found in (European Commission, 2016b). It is referred to as standardisation request M/530. However, we were unable to find the full text of the request.

[28] See a list of these standards at (ITEH Standards, n.d.).

expected to be cited in the OJEU for the GDPR[29].

While the GDPR is not based on compliance with standards, unlike the AI Act[30], it is nevertheless one of the first times that the Commission has requested standards for a European law in the digital field. Additionally, the EU considers privacy and data protection to be a fundamental right, enshrined in Article 8 of the Charter of Fundamental Rights (CFREU) (European Commission, 2012a). The standardisation request on privacy and personal data protection management was subsequently based on this same article (Kamara, 2017). The CEN-CENELEC JTC 13 standards are therefore the first attempt to develop standards relating to fundamental rights, something the AI Act aims to achieve on a larger scale[31].

3.4.4 *Voluntary certification in the GDPR*

Even if some technical standards exist for the GDPR, unlike the AI Act, the GDPR does not provide for the use of CE marking, as this mechanism is generally reserved for products covered by the NLF[32]. However, the GDPR uses voluntary certification, based on co-regulatory tools: codes of conduct and certification mechanisms, which play an important role in making controllers and processors accountable. These replace the traditional privacy seals[33], but are not published in the OJEU and do not benefit from the advantages of harmonised standards[34] and CE marking schemes.

Associations and bodies representing data controllers or processors in a given sector of activity can draw up codes of conduct to apply the GDPR rules to their sector or adapt them to the specific needs of enterprises (Art. 40(1) GDPR). Compliance with a code of conduct is voluntary, yet strongly encouraged, and is based on a self-certification mechanism with subsequent monitoring (Art. 41(1) GDPR). Other standardised mechanisms include binding corporate rules (Art. 47 GDPR) which allows companies to transfer personal data in and out of Europe, but that are mandatory once signed.

Furthermore, Article 42.1 provides for the *"establishment of data protection certification mechanisms and of data protection seals and marks"*. The purpose of these tools is to enable controllers and processors to demonstrate that their personal data processing complies with the Regulation. They can also be used to justify that a company that is not subject to the obligations of the GDPR complies with its principles and presents appropriate guarantees in the event of data being transferred outside the Union (Art. 42(2) GDPR). This is a voluntary process requiring an *a priori* assessment by an accredited certification body[35], a supervisory authority[36] or the European Data Protection Board (EDPB)[37]. If the EDPB considers criteria of a specific certification scheme consistent with the GDPR, this will result in a common certification called the "European Data Protection Seal" (Art. 42(5) GDPR).

In this context, the EDPB adopted in October 2022 an Opinion on the Europrivacy (n.d.a) criteria for certification (EDPB, 2022b). This marks the approval of the very first European Data Protection Seal (EDPB, 2022a). Companies can now evaluate their data processing against the Europrivacy criteria and show their compliance to the GDPR. With this status, Europrivacy certificates will be officially recognised in all EU countries. The Europrivacy certification covers the ISO/IEC 27001 standard (Europrivacy, n.d.b) which is already largely used, making the international criteria one of the components of this official European certification.

Certification thus plays a key role in the GDPR for the accountability of stakeholders, as adherence to approved codes of conduct or approved certification mech-

[29] See in particular the standard page on the CEN-CENELEC website: (CEN, n.d.b).

[30] For more information on the use of technical standards in the AI Act, see Chapter 4, Section 4.7.

[31] For more information on how standards can relate to fundamental rights, see Chapter 5.

[32] For more information on CE marking and the NLF, see Chapter 4, Section 4.6.

[33] A privacy seal is *"a certification mark or a guarantee issued by a certifying entity verifying an organisation's adherence to certain specified privacy standards that aim to promote consumer trust and confidence in e-commerce"* (Rodrigues et al., 2013).

[34] Harmonised standards are European standards that can be published in the OJEU and be granted certain legal properties. For more information, see Chapter 4, Section 4.6.2.

[35] Certification bodies issue and renew certification. They are accredited by either a supervisory authority or a national accreditation body (Art. 43(1) GDPR). To be accredited, a certification body must notably demonstrate its *"independence and expertise in relation to the subject-matter"* (Art. 43(2)(a) GDPR).

[36] A supervisory authority is an independent public authority established by a Member State (Art. 4(21) GDPR) and who is responsible for monitoring the application of the Regulation (Art. 51(1) GDPR).

[37] The EDPB is an independent Union body (Art. 68(1) GDPR) responsible for ensuring the consistent application of the Regulation (Art. 70(1) GDPR).

anisms may be used as an element by which to demonstrate compliance with the obligations of the controller and processor (Art. 24(3) and 28(5) GDPR). Moreover, multiple certification models co-exist in the GDPR and the text does not prohibit the establishment of certification schemes outside of Article 42 regime (Lachaud, 2020). However, some have argued that the certification process under the GDPR could already be seen as a new regulatory instrument (Lachaud, 2018), a form of self-regulation. Yet its scope remains limited and it should be seen as a stepping stone towards the extension of certification provided for in the AI Act.

It should be noted that contrary to the AI Act, certification in the GDPR is issued to a data controller or processor (Art. 42(7) GDPR), not to an infrastructure or product. Another difference is the body responsible for the certification assessment: while in the AI Act the conformity assessment is issued internally in some cases, in the GDPR, in the absence of hENs, the assessment will always be carried out by an external certification body. The main advantage of the certification mechanisms provided for by the GDPR is that they give a competitive advantage to companies that comply with them (Grafenstein, 2022). Despite these differences, the use of voluntary certification under the GDPR has therefore paved the way for prescriptive certification in the form of CE marking in the AI Act.

3.5 *Navigating the requirements of the AI Act*

3.5.1 *The scope of the AI Act*

The AI Act is an EU Regulation, which means it is directly applicable by Member States without the need to transpose it into national laws. The AI Act applies to various stakeholders across the EU and the European Economic Area (EEA)[38], including providers and deployers of AI systems. The provider is the entity responsible for developing the AI system (Art. 3(3) AI Act), while the deployer is the entity who uses the AI system (Art. 3(4)). AI system providers and deployers must ensure that their AI systems comply with the various requirements of the AI Act[39], depending on the category to which their system belongs[40]. Other stakeholders, such as distributors and importers, also have obligations when the AI system presents a high risk[41]. For instance, they must verify the various stages of the conformity assessment procedure (Art. 23 to 26).

The AI Act sets out two main purposes and two ways of achieving them. Indeed, the AI Act seeks both to *"improve the functioning of the internal market"* and to *"promote the uptake of human-centric and trustworthy AI"*. These objectives can be achieved by two main means: *"ensuring a high-level of protection of health, safety, fundamental rights"* and *"supporting innovation"* (Art. 1(1)). These two objectives are very distinct, and each of the provisions in the AI Act attempts to address one or the other. In particular, the text adopts an approach to product safety as a means of improving the internal market[42], while adapting it to the protection of fundamental rights[43]. The AI Act therefore conveys a dual discourse: improving the internal market and protecting fundamental rights.

To achieve these two objectives, the AI Act lays down various rules and requirements which apply in different contexts. However, for a system to be primarily covered by the AI Act, it must correspond to the material and territorial scope of the text. First, the definition of an AI system under the AI Act[44] covers most approaches known as "AI" in computer science[45]. Second, the AI Act applies specifi-

[38] In Europe, the twenty-seven countries of the European Union and the four countries of the European Free Trade Association (EFTA), namely Iceland, Liechtenstein, and Norway, are bound by the same rules governing the internal market and enabling the free movement of persons, goods, services, and capital within what is known as the "European single market".

[39] In particular with regard to Articles 16, 50, 53 and 55 of the IA Act for the provider, and Articles 26 and 50 for the deployer.

[40] See next paragraph for the different categories of AI systems.

[41] See the risk classification in Section 3.5.2.

[42] See Chapter 4, Section 4.7.

[43] See Chapter 5

[44] The definition is given in Chapter 1, Section 1.1.4.

[45] See Chapter 1, Section 1.1.3.

cally where a provider places an AI system on the EU market, whether or not it is located in the EU. It also applies when an AI system is used by a deployer whose registered office is located in the EU.

Stakeholders who fail to comply with the provisions of the AI Act may be subject to financial penalties. This fine can be as high as €35,000,000 or 7% of their total worldwide annual turnover for putting on the market a prohibited system. Other fines include, for example, €15,000,000 or 3% of the annual turnover for not complying with their obligations, and €7,500,000 or 1% of their worldwide annual if they fail to cooperate with national authorities.

3.5.2 *The risk based approach of the AI Act*

The AI Act is a risk-based regulation, meaning AI systems are classified into certain categories of risks and for each category, certain requirements apply. When the Commission released its first proposal, its representation of the risk hierarchy followed a pyramid shape with unacceptable risk systems at the top of the pyramid, followed by high risk, limited risk and, finally, minimal risk systems at the bottom, as shown in Figure 3.4.

As amendments have been made, the structure of these categories has evolved. Although the final version of the regulation retains its risk-based structure, the risk categories have changed and certain types of risk can be accumulated, making the pyramid representation obsolete. We propose an alternative representation in Figure 3.5. The two main categories of risk remain: unacceptable risk AI systems, and high-risk AI systems, respectively at the top of our representation.

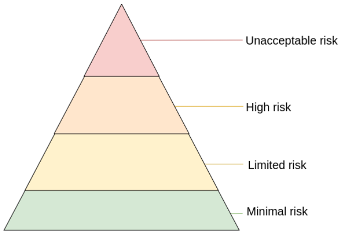


Figure 3.4: Pyramid of risks as initially intended by the Commission in the first draft of the AI Act. This representation, although still used by the European institutions, is now depreciated in the final version of the text. Image adapted from (European Commission, n.d.a).

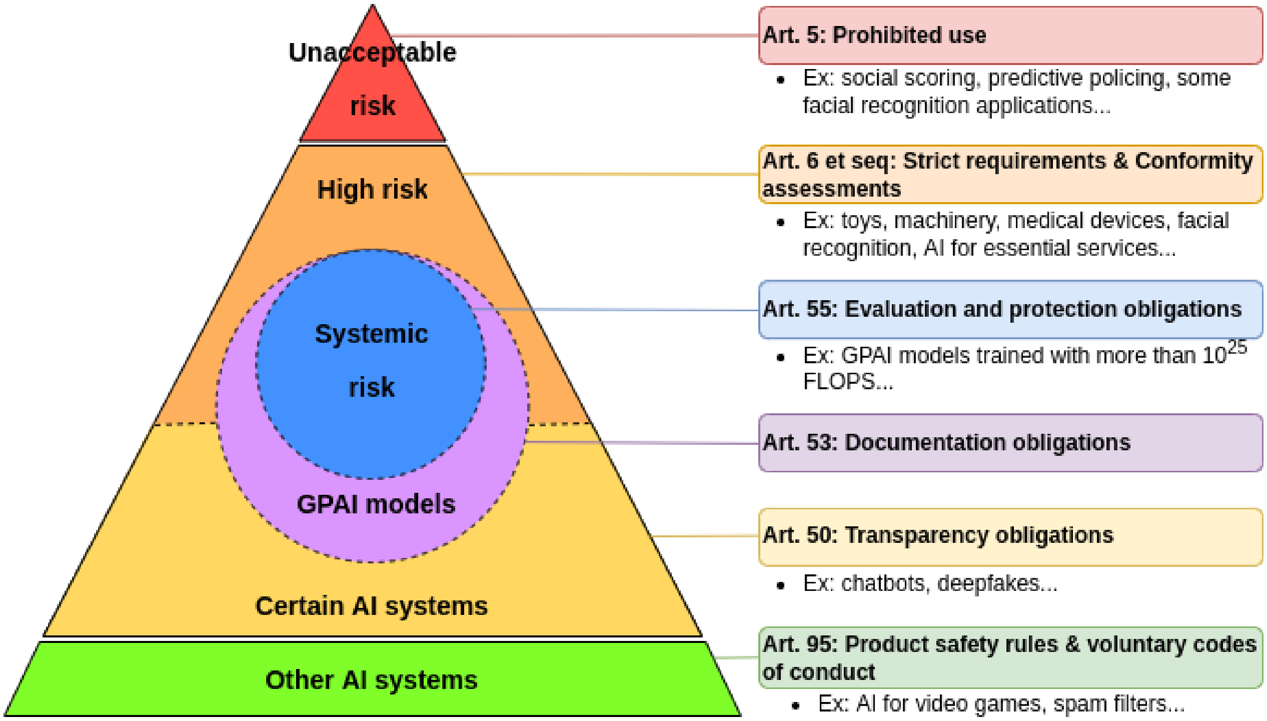

Unacceptable risk
High risk
Systemic risk
GPAI models
Certain AI systems
Other AI systems
Art. 5: Prohibited use
Ex: social scoring, predictive policing, some facial recognition applications...
Art. 6 et seq: Strict requirements & Conformity assessments
Ex: toys, machinery, medical devices, facial recognition, AI for essential services...
Art. 55: Evaluation and protection obligations
Ex: GPAI models trained with more than 10^{25} FLOPS...
Art. 53: Documentation obligations
Art. 50: Transparency obligations
Ex: chatbots, deepfakes...
Art. 95: Product safety rules & voluntary codes of conduct
Ex: AI for video games, spam filters...

Figure 3.5: Visual representation of the categories of AI systems in the AI Act (by the author). Dotted lines represent categories which can be accumulated.

Unacceptable risk

The use of AI systems which present an unacceptable risk is totally prohibited by the AI Act. For these systems, the EU has opted for a precautionary approach (Almada and Petit, 2023). Unacceptable risk systems include – but are not limited to – social scoring[46], predictive policing[47], emotion recognition in the workplace or the education system[48], biometric classification[49] and biometric identification, including facial recognition, under certain conditions[50] (Art. 5(1)).

High risk

There are two ways to fall into the high-risk category: either (i) the product using AI is already covered by EU harmonised legislation; or (ii) the domain of application of the AI system must be listed in Annex III (Art. 6). The EU harmonised legislation for (i) is listed in Annex I. It contains the twelve NLF regulations, such as those on machinery, toys, lifts, radio equipment and medical devices, as well as other legislation covering, for instance, certain motor vehicles or aircraft. The AI system must be either the safety component of a product covered by one of the regulations, or itself a product covered by the regulation. In addition, it must undergo a conformity assessment by a third party in accordance with this regulation (Art. 6(1)). In addition, for (ii), systems covered by Annex III, called *"stand-alone AI systems"* (rec. 52) include biometric systems, AI systems used for critical infrastructure, education, employment, essential public services, law enforcement, migration and justice. There are, however, exceptions whereby systems listed in Annex III may not be considered high risk, for instance if they are intended to perform a *"narrow procedural task"* or to simply improve the result of a human activity (Art. 6(3)). All AI systems listed in Annex III will be registered in an EU database (Art. 71(1)). The high-risk category should cover approximately 5 to 15% of all AI systems in the EU, according to the European Commission (2021b). However, other studies show that this figure could actually be much higher (appliedAI, 2023).

High risk AI systems have to comply with a list of essential requirements that are detailed in Chapter III, Section 2 and 3 of the AI Act[51]. Under these requirements, the provider should: establish a risk management system and a quality management system (Art. 9 and 17), use quality data (Art. 10), draw up technical documentation and make it available to national authorities (Art. 11 and 18), record events in logs and keep the logs for an appropriate period of time (Art. 12 and 19), ensure a level of transparency which enables output interpretation (Art. 13), design systems to be overseen by humans (Art. 14), ensure accuracy, robustness and cybersecurity of the system (Art. 15). In addition, public entities or private bodies providing public services must carry out an assessment of the AI system's impact on fundamental rights (Art. 27).

To be distributed on the EU market, high risk AI system must undergo a conformity assessment procedure and receive a European Conformity (CE) mark[52] which shows compliance with the regulation. This procedure may be carried out by a third party or be a self-assessment carried out directly by the company, depending on the application of the AI system. To demonstrate compliance, providers will rely in particular on the state of the art and on harmonised technical standards[53].

[46] Systems which evaluate the social behavior of individuals.

[47] Systems which predict the risk of a person to commit a criminal offence.

[48] Unless for medical or safety reasons.

[49] Systems designed to deduce protected characteristics of individuals, such as race, political opinion, religious belief, sexual orientation and so on, by using their biometric data.

[50] When it is used in real-time and remotely, in a public space, for law enforcement purposes, and when it does not fall under a list of exceptions such as the targeted search for victims of human trafficking or perpetrators of serious criminal offences, or the prevention of terrorist attacks.

[51] Chapter III, Section 2, related to the obligations of the providers, while Section 3 also encompasses obligations for deployers and other parties.

[52] For more information on CE marking in the EU, see Chapter 4, Section 4.6.4.

[53] For a discussion on the conformity assessment process, harmonised standards and CE marking in the AI Act, see Chapter 4, Section 4.7.

Certain AI systems

The limited risk category first proposed by the European Commission is replaced by the "certain AI systems" category in the final version of the text. This category is cumulative with the category of high-risk systems. This change to the pyramid structure is due mainly to the inclusion in the text of General Purpose AI (GPAI) systems, capable of performing a wide range of tasks[54], and which, depending on their application, may or may not fall into the high-risk category.

The "certain AI systems" category thus includes GenAI systems as a whole[55] – including GPAI systems; emotion recognition or biometric categorisation systems whose application does not fall under the high-risk classification; and AI systems in general which are *"intended to interact with natural persons"*. These systems are subject to transparency obligations. Providers are required to inform individuals when they interact with an AI system and deployers of emotion recognition or biometric categorisation systems must inform individuals when they are subject to the operation of these systems. For GenAI, providers must ensure that the results of their GenAI system can be detected as artificially generated and GenAI deployers must disclose that content has been artificially generated.

GPAI models

GPAI models now have their own category which is also cumulative with others. Indeed, GPAI models can be integrated into an AI system and thus fall into, at least, the "certain AI systems" category and, for certain applications, may also fall into the "high risk" category. GPAI models, regardless of the other risk categories in which they may be classified – certain AI systems and/or high-risk – have additional obligations. Providers must: (i) draw up technical documentation, including on the training and testing processes and results, and make it available upon request to the AI Office; (ii) provide information and documentation to providers who intend to integrate the GPAI model into their AI system; (iii) put in place a policy on copyright; and (iv) make publicly available a detailed summary about the content used for training. To show compliance with these requirements, providers can rely on codes of practices approved by the Commission, or on harmonised standards if they exist (Art. 53(4)).

Systemic risk

In addition to the requirements specific to GPAI models and the requirements linked to other risk categories – certain AI systems and/or high risk – GPAI models which present "systemic risks", due to their scale and the importance of their potential impact, are subject to additional requirements. This is notably the case of models which use an amount of computation of more than 10^{25} floating point operations (FLOP)[56] for training. However, if any models above this threshold is automatically considered with systemic risk, it is a necessary condition to fall in this category. Indeed, high impact capabilities can also be evaluated on the basis of *"appropriate technical tools and methodologies"*. The Commission might decide, following a alert from the scientific panel that a GPAI model presents a systemic risk. This can be assessed using the criteria in Annex XIII, such as the number of model parameters, the quality or size of the dataset and the amount of computation. The

[54] GPAI *systems* are based on GPAI *models*. The complete definition of GPAI models is given in Chapter 1, Section 1.1.4.

[55] The term GenAI is not directly used by the AI Act, which refers to AI systems *"generating synthetic audio, image, video or text content"* (Art. 50(2)).

[56] A FLOP is a simple mathematical operation, such as addition or division, performed with floating-point numbers, which are approximations of decimal numbers.

type of input and output modality should be taken into account and compared with the state of the art in the field[57]. The Commission could also examine the impact on the internal market and the number of end-users.

Additional requirements for GPAI models with systemic risk fall on the provider, who must evaluate the model using standardised protocols, including adversarial testing, mitigate risk, keep track of and report serious incidents to the AI Office, and ensure cybersecurity protection. To show compliance with these requirements, as with the other GPAI models, providers of GPAI models with systemic risk can rely on codes of practice or harmonised standards (Art. 55(2)).

Ultimately, some AI systems will fall into up to four categories, each with its own specific requirements. This is the case for GPAI models (GPAI models and certain AI systems category), with high capabilities (systemic risk category), which are used for high risk applications (high risk category). It is worth noting, however, the difference between the GPAI model and the AI system in which it is integrated downstream. The GPAI model and the downstream high-risk system may have different providers, each with their own obligations.

Other AI systems

There is no official name for the category of AI systems which are neither unacceptable risk, high-risk, nor GPAI systems. Indeed, other AI systems are not subject to any specific requirements under the AI Act. However, they are mentioned in the text, which notably encourages providers and developers to draw up of voluntary code of conduct (Art. 95). In addition, it should be noted that all AI systems fall under the General Product Safety Regulation (European Parliament and Council, 2023c). Recital 166 describes it as a "safety net" for systems which are not considered high risk. In addition, AI systems that are products in areas already covered by harmonised legislation will have to follow these sector-specific regulations.

3.5.3 *Testing without hindering innovation*

Alongside obligations, the AI Act also includes measures *"in support of innovation"* (Chapter VI). These measures include AI regulatory sandboxes. The AI Act defines a sandbox as *"a controlled environment that fosters innovation and facilitates the development, training, testing and validation of innovative AI systems for a limited time before their being placed on the market or put into service"* (Art. 57(5)). Each member state must establish at least one regulatory AI sandbox. Sandboxes serve both to empower businesses to innovate and to foster compliance by allowing stakeholders to learn about regulatory obligations (Art. 57(9)). This will be particularly useful for SMEs, which have priority access to sandboxes (Art. 62(1)(a)). In a sandbox environment, providers remain liable for damages, but will not be prosecuted for breaching the AI Act (Art. 57(12)).

The EU Commission will provide technical support and advice on the establishment and operation of sandboxes. In particular, the AI sandboxes will submit an annual report to the Commission[58]. Further details on the operation of AI sandboxes will be provided in Commission implementing acts (Art. 58(1)).

But real-world testing can also take place outside AI sandboxes, particularly when testing high-risk AI systems (Art. 60(1)). These testings can only take place after a real-world testing plan has been approved by market surveillance authori-

[57] The state of the art, for example in FLOP, can vary between text and image generation.

[58] Specifically to the AI Office and the AI Board. See below for more information on these entities.

ties[59] (Art. 60(4)) and after the consent of the subjects of testing has been obtained prior to their participation (Art. 61(1)).

3.5.4 *A new governance ecosystem*

To implement and enforce requirements, the AI Act rely on the market surveillance scheme within the meaning of Regulation 2019/1020 (European Parliament and Council, 2019), but also create a brand new ecosystem. As such, the AI Act provides for new bodies, to ensure that the law is properly implemented. Together with existing entities, they should create a European AI governance ecosystem and make sure that legal requirements are met. This choice is justified by recital 148 which stipulates that the governance framework should allow to *"coordinate and support the application of this Regulation at national level, as well as build capabilities at Union level and integrate stakeholders in the field of AI"*. This ecosystem is represented in Figure 3.6.

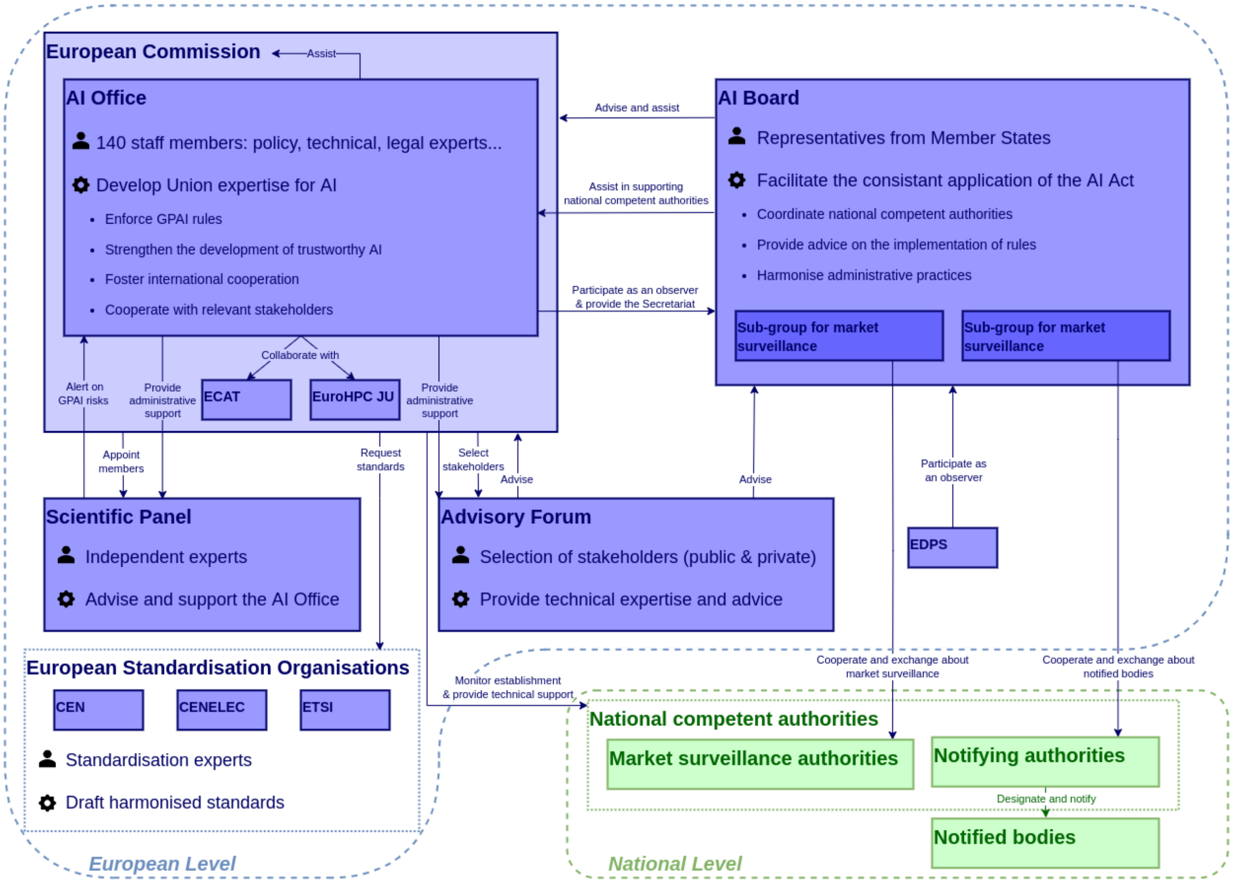

European Commission
Assist
AI Office
140 staff members: policy, technical, legal experts...
Develop Union expertise for AI
Enforce GPAI rules
Strengthen the development of trustworthy AI
Foster international cooperation
Cooperate with relevant stakeholders
Advise and assist
Assist in supporting national competent authorities
Participate as an observer & provide the Secretariat
AI Board
Representatives from Member States
Facilitate the consistant application of the AI Act
Coordinate national competent authorities
Provide advice on the implementation of rules
Harmonise administrative practices
Sub-group for market surveillance
Sub-group for market surveillance
Collaborate with
Alert on GPAI risks
Provide administrative support
ECAT
EuroHPC JU
Provide administrative support
Appoint members
Request standards
Select stakeholders
Advise
Advise
Participate as an observer
Scientific Panel
Independent experts
Advise and support the AI Office
Advisory Forum
Selection of stakeholders (public & private)
Provide technical expertise and advice
EDPS
European Standardisation Organisations
CEN
CENELEC
ETSI
Standardisation experts
Draft harmonised standards
Monitor establishment & provide technical support
Cooperate and exchange about market surveillance
Cooperate and exchange about notified bodies
National competent authorities
Market surveillance authorities
Notifying authorities
Designate and notify
Notified bodies
European Level
National Level

Figure 3.6: Representation of the European AI governance ecosystem.

National governance

On national level, national competent authorities, consisting of at least one market surveillance authority and one notifying authority, must be designated by each Member State (Art. 70(1) AI Act).

Market surveillance authorities are in charge of supervising the placing on the market of AI systems. They act as a point of contact and interface between the public and the Union. For systems covered by existing NLF legislation, existing author-

[59] See next paragraph for more information on these entities.

ities are automatically designated to also monitor AI systems in their sector. They lead all investigation operations in their sector, with the only exception of GPAI models, which are mainly monitored by EU governance bodies. For these investigations, market surveillance authorities may be granted access to the source code under certain conditions (Art. 74(12)). Together with the Commission, these market surveillance authorities can also propose *"joint investigations"* to promote compliance or identify non-compliance (Art. 74(11)). Finally, they can require suppliers to take corrective action in the event of regulatory non-compliance and take action if suppliers refuse to comply (Art. 79(5)). The AI Act also provides a mechanism for any natural or legal person to lodge a complaint with their market surveillance authority if they believe there has been an infringement of the AI Act (Art. 85).

On the other hand, notifying authorities designate and notify conformity assessment bodies to become *"notified body"*, as well as monitor their activities. Notified bodies are responsible for carrying out mandatory conformity assessments for AI systems that require them. They are at the center of the compliance assessment mechanism for high-risk AI systems put in place by the AI Act[60].

AI Office

At European level, the first of the new entities created by the AI Act is the AI Office, set up within the European Commission[61] to *"develop Union expertise on AI"* (Art. 64(1) AI Act). The AI Office was established by the European Commission Decision of 24 January 2024 (European Commission, 2024c), hereafter EC Dec. The AI Office is notably tasks to enforce the rules on GPAI (Art. 2(1) EC Dec) but it also has other missions, such as encouraging innovation or fostering cooperation at different levels (Art. 2(2) EC Dec).

To ensure compliance with GPAI rules, the AI Office will be specifically tasked with developing tools and benchmarks for evaluating GPAI models. It will also have a monitoring role, particularly with regard to the emergence of new risks, and the correct implementation of GPAI requirements. For instance, it will conduct evaluation of GPAI models and investigate potential infringements on GPAI rules (Art. 3(1) EC Dec). But the AI Office will also be responsible for drafting codes of practice for GPAI models, technical specifications that will help stakeholders comply with the AI Act's GPAI requirements, pending the development of harmonised standards (Art. 56(1) AI Act). Providers of GPAI models will be invited to participate in the drawing-up of codes of practice (Art. 56(3) AI Act).

The AI Office will also play an important role in assisting the EU Commission, by preparing Commission Decisions, implementing acts and delegated acts. It will also oversee the proper development of standards, prepare standardisation requests and common specifications if necessary[62]. Finally, it will prepare guidance and guidelines in support of the AI Act and provide advice on the implementation of AI sandboxes and real world testing with national competent authorities (Art. 3(2) EC Dec). The AI Office will also keep a list of planned and existing AI sandboxes (Art. 57(15) AI Act).

Finally, the AI Office will cooperate with other entities: firstly, within the Commission by working with other services (Art. 5 EC Dec)[63]; secondly, at international level by supporting other similar institutions or agreements (Art. 7 EC Dec); and thirdly, with expert stakeholders, including the industry (Art. 4 EC Dec). This last point includes overseeing the AI Pact, an initiative to promote the industry's volun-

[60] For more information on conformity assessment procedures for AI systems, see Chapter 7.

[61] More specifically, within the Directorate-General for Communication Networks, Content and Technology (DG CONNECT).

[62] Common specifications are a fallback solution in case traditional standards fail. They can only be drafted under certain conditions (Art. 41). For more information, see Chapter 7, Section 7.9.4.

[63] Notably the European Centre for Algorithmic Transparency (ECAT) in charge of developing Union expertise for large online platforms and enforce the DSA (Art. 5(2)(a) EC Dec); or the European High Performance Computing Joint Undertaking (EuroHPC JU), an initiative with private actors to develop a supercomputing ecosystem in Europe (Art. 2(3)(c) EC Dec).

tary commitment to the AI Act requirements ahead of the legal deadline in order to anticipate its impact (European Commission, n.d.b).

As regard to its inner structure, the AI Office will be separated into five units: "Excellence in AI and robotics", "AI Regulation and Compliance", "AI Safety", "AI Innovation and Policy Coordination", and "AI for Societal Good", called CONNECT.A.1 to CONNECT.A.5. They will be supported by a "Lead Scientific Advisor" and an "Advisor for International Affairs". This structure is presented in Figure 3.7. The structure of the AI Office is in fact a reorganisation of Unit A of the European Commission's Directorate General of Communications Networks, Content and Technology (DG CONNECT). The changes between DG CONNECT A and the AI Office will not require a *"huge reorganisation"*, but the task force will be improved as the AI Office plans to recruit more than 80 people over the next two years, bringing the total workforce over 140 (Gkritsi, 2024).

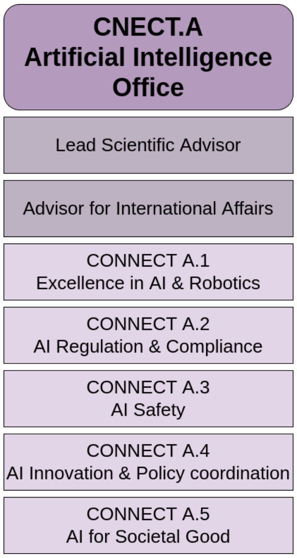

CNECT.A
Artificial Intelligence Office
Lead Scientific Advisor
Advisor for International Affairs
CONNECT A.1
Excellence in AI & Robotics
CONNECT A.2
AI Regulation & Compliance
CONNECT A.3
AI Safety
CONNECT A.4
AI Innovation & Policy coordination
CONNECT A.5
AI for Societal Good

Figure 3.7: Structure of the AI Office, adapted from (European Commission, n.d.f). All units are independent, there is no hierarchy between units.

AI Board

But the AI Office will not be alone in monitoring the correct implementation of the AI Act. The text also provides for the creation of an AI Board, hereafter "the Board", composed of one representative per Member States. The European Data Protection Supervisor will also participate as an observer (Art. 65(2) AI Act). The AI Office will attend the Board's meetings without taking parts in the votes (Art. 65(2)) and provide Secretariat for the Board (Art. 65(8)). The Board is tasked with supervising that the AI Act is applied consistently in all Member States. This includes coordinating national authorities, providing advice on the implementation of rules and monitor the harmonising of practices (Art. 66(a),(c),(d)). The Board will also issue recommendations at the request of the Commission, in particular on existing standards and their use, as well as on common specifications (Art. 66(e)(iii),(iv)).

In practice, a large part of the Board's work will take place in thematic subgroups. The AI Act provides for the creation of a subgroup on market surveillance and another on notified bodies, but members of the Board may suggest the creation of new subgroups (Art. 65(6)). Recently, journalists reported that other subgroups were already in the making, in particular on technical standards, GPAI, innovation and regulatory sandboxes, prohibited systems, high-risk categorisation, or on the interplay with other EU legislation (Bertuzzi, 2024).

Advisory Forum and Scientific Panel

The AI Act also establishes two auxiliary entities: (i) the Advisory Forum, composed of a selection of stakeholders; and (ii) the Scientific Panel, composed of independent experts. The stakeholders of the Advisory Forum and the experts of the Scientific Panel are both appointed by the Commission, but while the experts of the Scientific Panel should be independent from any AI system provider, the Advisory Forum is composed of a wide range of stakeholders, including industry, Small and Medium-sized Enterprises (SMEs), startups, academia and civil society (Art. 67(2)). Relevant EU agencies and European Standardisation Organisations are also members of the Advisory forum (Art. 67(5)).

The purpose of the Advisory Forum is to provide technical expertise and advice to the Board and the Commission. For instance, the Commission will consult the Advisory Forum before drafting standardisation requests (Art. 40(2)) and common specifications (Art. 41(1)(b)§2). The Scientific Panel is tasked with working with the AI Office, raising alert on GPAI models that could be considered with systemic risk (Art. 90), helping with the drafting of codes of conducts and supporting the monitoring activities of the AI Office. Experts from the scientific groups may be called upon by the Commission to carry out the evaluation of GPAI models (Art. 92(2)). They are also available to Member States requiring expert advice (Art. 69).

The objectives of the two bodies are therefore strictly different: whereas the Advisory Forum is designed as a platform for stakeholders to express their interests and raise concerns about the implementation of sector-specific measures, the Scientific Panel is supposed to be independent, impartial and objective (Art. 68(4)).

3.5.5 *Various deliverables are meant to support the legal requirements*

The EU governance ecosystem put in place for AI has an important role to play in the proper implementation of the AI Act and various entities will be involved in developing deliverables which will help support the legal text.

Harmonised standards, developed by the European standardisation bodies at the request of the European Commission, will be the most important of these deliverables. They are expected to define the technical requirements for specifying the legal obligations of the AI Act[64]. They will have direct legal effects and, as such, will be the preferred means of compliance with the requirements set out by the AI Act for high-risk AI systems and GPAI models. As of now, they are mainly drafted for high-risk AI systems, but the Commission may also request standards for GPAI models in the near future.

For the other obligations set out in the AI Act, the European institutions have made provisions for other means of compliance, albeit less powerful than harmonised standards. In particular, codes of practice are supposed to provide elements of compliance for GPAI models providers (Art. 53(4)). They do not benefit from the advantages of harmonised standards but may nevertheless be approved by the European Commission by means of implementing acts, in order to give them general validity in the EU (rec. 117). However, they are only intended to supplement the absence of harmonised standards, which would take precedence once published (Art. 55(2)). Codes of practice will be drawn up by a group of relevant stakeholders, with the assistance of the AI Office and the support of the scientific panel. The AI Office and the AI Board will monitor their implementation (Art. 56(1)). The stakeholders responsible for drafting these codes of practice should include, in particular, providers of GPAI models and national competent authorities (Art. 56(3)). Four working groups are currently responsible for drafting the GPAI codes of practice[65], with chairs from different areas of expertise (European Commission, 2024f), selected by the AI Office. After a multi-stakeholders consultation on the codes of practice in the summer of 2024 (European Commission, 2024b), the first meeting of the working groups took place on 30 September 2024, with about a thousand participants (European Commission, 2024e). Discussions are expected to last until 30 April 2025, when a final draft should be presented[66] (European Commission, 2024f).

Codes of conduct are another deliverable that will complement the AI Act. Codes of conduct are voluntary frameworks adopted by providers of non-high-risk AI systems and encouraged by Member States to advance AI literacy (rec. 20). Their development is less stringent and they could emanate from different stakeholders,

[64] For more information on harmonised standards, see Chapter 4, Section 4.6.2.

[65] These are "Transparency and copyrighted-related rules", "Risk identification and assessment, including evaluations", "Technical risk mitigation", "Internal risk management and governance of general-purpose AI providers"

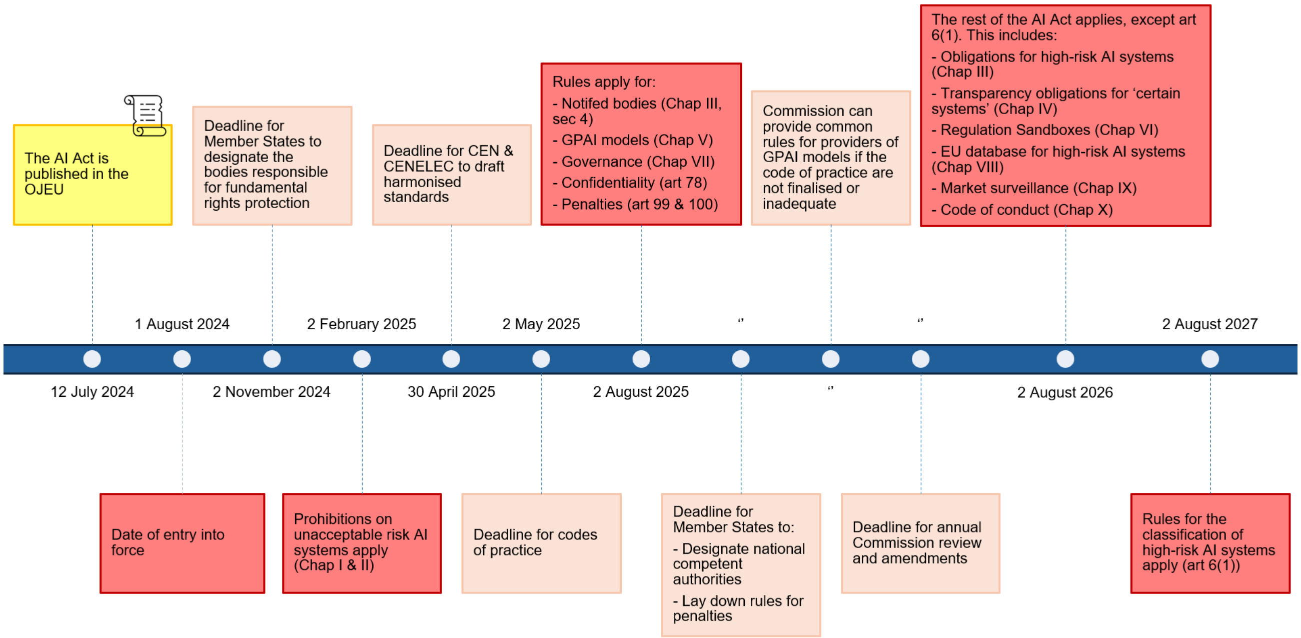
[66] The strict deadline imposed by the AI Act is 2 May 2025 (Art. 53(9)).

including providers or deployers of AI systems, although the Commission may also contribute to the development of such initiatives, in particular through the AI Office (Art. 95(3)). However, the AI Act specifies that such codes should nevertheless be *"developed in an inclusive way"*, with the help of relevant stakeholders, including from civil society and academia (rec. 165).

In addition to codes of practice and codes of conduct, the Commission is empowered to adopt delegated and implementing acts, and to adopt guidelines, with the supervision of the AI Board. Delegated acts will mainly modify requirements of the AI Act, for example by amending the annexes or the conditions for a system to fall into a specific category. Guidelines, on the other hand, will clarify certain requirements of the AI Act, such as when a system should be considered in a certain risk category or how to apply certain provisions of the text. Finally, implementing acts will allow for the approval of existing frameworks or frameworks developed by the Commission itself. In particular, they will be used to approve a code of practice for the transparency obligations of certain AI systems (Art. 50(7)) and for the obligation of GPAI models (Art. 56(6)). They will also be used to establish "common specifications" in the absence of adequate harmonised standards[67] (Art. 41(1)) and "common rules" in the absence of adequate codes of conduct (Art. 50(7) and 56(9)). We provide an organised list of what delegated acts, guidelines and implementing acts may contain in Appendix A.2.

3.5.6 *Entry into force*

After its publication in the OJEU, the text will now be implemented through several steps, represented in Figure 3.8.



12 July 2024

The AI Act is published in the OJEU

1 August 2024

Date of entry into force

2 November 2024

Deadline for Member States to designate the bodies responsible for fundamental rights protection

2 February 2025

Prohibitions on unacceptable risk AI systems apply (Chap I & II)

30 April 2025

Deadline for CEN & CENELEC to draft harmonised standards

2 May 2025

Deadline for codes of practice

2 August 2025

Rules apply for:
- Notifed bodies (Chap III, sec 4)
- GPAI models (Chap V)
- Governance (Chap VII)
- Confidentiality (art 78)
- Penalties (art 99 & 100)

‘‘

Deadline for Member States to:
- Designate national competent authorities
- Lay down rules for penalties

‘‘

Commission can provide common rules for providers of GPAI models if the code of practice are not finalised or inadequate

‘‘

Deadline for annual Commission review and amendments

2 August 2026

The rest of the AI Act applies, except art 6(1). This includes:
- Obligations for high-risk AI systems (Chap III)
- Transparency obligations for ‘certain systems’ (Chap IV)
- Regulation Sandboxes (Chap VI)
- EU database for high-risk AI systems (Chap VIII)
- Market surveillance (Chap IX)
- Code of conduct (Chap X)

2 August 2027

Rules for the classification of high-risk AI systems apply (art 6(1))

Figure 3.8: Timeline of implementation of the AI Act.

Indeed, the AI Act first came into force on 1 August 2024 but many requirements will apply later. Prohibitions on unacceptable risk systems will apply from 2 February 2025, obligations for GPAI models will apply from 2 August 2025 and transparency obligations from certain AI systems will apply from 2 August 2026.

[67] For more information on the condition of establishment of common specifications, see Chapter 7, Section 7.9.4. For the opinion of standardisation experts on common specifications, see Chapter 8, Section 8.5.4. For a discussion on what the duplication of frameworks under the AI Act might reveal about the lack of trust in standardisation, see Chapter 9, Section 9.5.1.

For high-risk systems, requirements will apply from 2 August 2026 for systems listed in Annex III but only from 2 August 2027 for systems covered by harmonised legislation listed in Annex I.

3.6 *Criticising the AI Act: what scope, obligations and enforcement mechanisms?*

When the Commission first proposed the draft AI Act in 2021, the text attracted a large number of criticisms. Some of them have been addressed in the final version, but the most structural criticisms remain unchanged. In this section, we briefly go through these criticisms[68].

3.6.1 *A complicated agreement on the definition of AI systems*

The definition of an AI system has evolved significantly since the European Commission's first proposal in April 2021. Notably, in the Commission's proposal (European Commission, 2021e), the definition of AI systems was left essentially to Annex I, which, at the time, contained a list of three types of approaches which could be considered AI: (a) Machine learning approaches, (b) Logic- and knowledge-based approaches, and (c) Statistical approaches and optimisation methods.

The first two approaches refer to the two main families of AI[69], while the third encompasses certain computer programs not normally considered AI. Scholars have pointed out that it was too broad a definition, likely to give rise to legal uncertainty (Ruschemeier, 2023). Some even argued that only machine learning systems should be regulated by the AI Act. In their views, this broad scope was justified for AI systems that present unacceptable risks, as the ban of these systems is justified by the dangers they pose to society and individuals regardless of the technology utilised. However, these critics believed that obligations for high-risk systems were based on characteristics specific to ML systems, such as opacity or dependency on data (Ebers et al., 2021) and that encompassing other systems in the scope of these obligations would lead to overregulation.

In the final version of the text, the definition adopted[70] does not refer directly to technical terms, and emphasises instead the autonomy of these systems. This definition is much closer to the definition given by the Organisation for Economic Co-operation and Development (OECD)[71]. According to some scholars, this definition further broadens the scope of the AI Act, moving towards a definition of software rather than AI (Hacker, 2024). The clarifications provided by the various recitals draw a clearer line between simple software and AI, but they still leave gaps, as it is not clear whether statistical approaches are taken into account. It will be up to case law to define the frontiers of what is considered AI and what is not.

3.6.2 *What systems should be regulated?*

Criticism over the AI Act focuses primarily on the scope of the legal text, i.e. the exact nature of the systems regulated. According to Smuha et al. (2021) the list of prohibited systems is too restrictive. For instance, military applications are excluded from the AI Act. For some scholars, this is a significant gap (Smuha et al., 2021), as AI applications for defence purposes raise many ethical and deontological questions, especially as many scholars are now calling for a ban on

[68] It should be noted that this list of criticisms is not exhaustive. Interested readers can consult the references cited, which generally provide other points of criticism in addition to those mentioned here.

[69] See Chapter 1, Section 1.1.3.

[70] See Chapter 1, Section 1.1.4.

[71] In its "Recommendation on AI" first adopted in 2019, the OECD defines an AI system as *"a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy"* (OECD, 2019). This definition was slightly modified in 2024 to get even closer to the definition of the AI Act (OECD, 2024b).

autonomous weapons systems (Brand, 2022). This omission could also pose problems for systems that have a dual use and can be used for both civil and military purposes (Ruschemeier, 2023). In addition to military and defence purposes, non-professional purposes and systems used solely for research purposes[72] are also excluded from the scope of the AI Act, leaving the door open to potentially harmful systems (Smuha and Yeung, 2024).

Even in systems that are actually included in the list of prohibitions, certain questions remain. The inclusion of subliminal techniques, for example, has left some researchers wondering what it could possibly contain (Ebers et al., 2021). Furthermore, the ban on social scoring in the AI Act is limited to public entities, ignoring the use of such technologies in the private sector, with dangerous applications that are not prohibited by the AI Act, such as credit scoring (Ebers et al., 2021).

One of the most debated application was probably biometrics. Indeed, the story of the inclusion of biometrics in one category or the other has been one of back and forth. Since the amendments from the three European institutions contained various exceptions and inclusions, biometrics and facial recognition in particular was a sensitive topic during the trilogue negotiations (Bertuzzi, 2022a), alongside military applications. In particular, some countries, such as France, strongly opposed the inclusion of military applications and pushed for more exceptions on facial recognition. The adopted version finally comes close to the first proposal of the Commission, with a ban on biometrics systems used by law enforcement authorities in specific circumstances[73]. This ban is further accompanied by broad exceptions which, according to some scholars, do not protect individuals against the dangers of these technologies for fundamental rights (Ruschemeier, 2023). Some have proposed to extend the ban to any biometric system used in public spaces (Ebers et al., 2021), or have called for an additional ban on any emotion recognition system (Wachter, 2024), without success.

The list of high-risk AI systems was also criticised, with some scholars considering that it was too restrictive (Smuha et al., 2021) and suggesting applications that could be added, such as AI systems for housing purposes (Ebers et al., 2021). For Edwards (2022), although the Commission retains the right to modify this list in theory, in practice it will probably be difficult to add new systems to the list. Furthermore, the negotiations during the trilogue resulted in the addition of some exceptions to the classification of high-risk AI systems, such as when a system is intended to perform a narrow task or simply complement human activity without replacing it (Art. 6(3)). This last-minute change to the text has been heavily criticised, as it complicates the assessment of a system's risk category and creates dangerous loopholes (Wachter, 2024).

Finally, in previous versions of the text, the question of open-source AI system was not addressed, leading scholars to wonder whether the obligations will be the same as for other systems (Ebers et al., 2021). The final version of the AI Act clarifies that the text does not apply to open-source systems unless they fall into the category of unacceptable risk, high risk or GPAI (Art. 2(12)). This broad exception leaves many open-source system applications still regulated. However, providers of open-source high-risk AI systems and of open-source GPAI models which do not present a systemic risk are exempt of some information and documentation obligations (Art. 25(4) and Art. 53(2)).

[72] On the contrary, some consider that the first version of the AI Act lacked exceptions for research purposes (Ruschemeier, 2023; Ebers et al., 2021). This problem has therefore been resolved in the current final version, which explicitly excludes them.

[73] In public spaces, in real-time.

3.6.3 *What should be required of AI systems?*

The requirements for high-risk AI systems were also at the center of debates, with some scholars considering that the provisions not sufficient to protect against the harms generated by AI and ensure protection of fundamental rights (Smuha et al., 2021). For instance, the data governance requirements fail to explain which biases should be mitigated and what types of discrimination are considered (Ebers et al., 2021), as well as how these biases could be mitigated, while the academic literature flourishes on these issues (Wachter, 2024). Wachter (2024) also highlights the absurdity of requiring a dataset to be "representative" in a world where any set of historical data is biased, explaining that *"neutral data is a fantasy"*. As Article 10 is included in one of the Standardisation Request items, these points should be subsequently covered by harmonised standards[74].

Other requirements of high-risk AI systems were also criticised. For instance, some consider the requirement on human oversight to be impractical, as it is not yet possible to fully understand a system as the article would require, and the AI Act does not specify when oversight is necessary (Ebers et al., 2021).

Furthermore, the transparency requirement does not say any thing on the interpretability of the systems' output (Ebers et al., 2021), although it should be noted that a "right to explanation of individual decision-making" has been added in the final version (Art. 86). It provides for the right to obtain from the deployer *"clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken"*. However, this right is only applicable to the high-risk systems listed under Annex III point 1 and point 3 to 8. It is not applicable for AI systems used in critical infrastructures (Annex III point 2), nor high-risk AI systems that already fall under product safety regulations listed in Annex II.

In addition, some worry about the newly added obligation to conduct a FRIA (Art. 27). FRIAs were added thanks to the (European Parliament, 2023a, amendment 413), after their absence in the Commission's version of the AI Act was heavily criticised (Edwards, 2022). However, for Hacker (2024), FRIAs are unlikely to be effective, as they are simply a means to tick boxes, rather than genuinely assess fundamental rights. Some also point out that FRIA are only an obligation of deployers and not of providers of high-risk AI systems, that they only apply to deployers that are public entities, acting on behalf of public entities or providing public services (Wachter, 2024). Furthermore, as with the right to an explanation, the provision to conduct a FRIA excludes critical infrastructures and AI systems that already fall under product safety regulations.

Finally, the absence of certain requirements is also criticised, as the environmental cost of these systems is hardly taken into account (Wachter, 2024; Hacker, 2023).

In addition to facial recognition and military applications, discussions during the trilogue were also heated regarding Generative AI (GenAI) (Bertuzzi, 2023). According to some scholars and journalists, the provisions relating to GPAI models are the result of strong lobbying efforts by GenAI companies such as the French company Mistral AI or the German company Aleph Alpha (Wachter, 2024; Chan, 2023). As a result, the obligations of GPAI models focus more on transparency than on liability, and compliance is assessed through codes of practices rather than hard regulation. These rules on GPAI are seen as *"extremely weak"* by some scholars (Hacker, 2023), who are therefore calling for additional obligations to be added, such as guaranteeing a high level of cybersecurity for GenAI models (Hacker, 2024).

[74] For more information about Article 10 and standards on fairness and data quality, see Chapter 6.

In addition, the 10^{25} FLOPS threshold is likely to cover very few current systems, since the freely available version of ChatGPT and the Mistral and Aleph models are, for instance, below this threshold (Wachter, 2024). Scholars also criticise the choice of a strict threshold which does not necessarily represent a level of danger and remains highly arbitrary (Smuha and Yeung, 2024). Such threshold may encourage providers of GPAI models to remain below the threshold without reducing the harmful effects of their models (Wachter, 2024).

3.6.4 The use of CE marking

Some scholars have argued that making the AI Act a product safety regulation is a way of playing on the EU's strength (Almada and Petit, 2023). Indeed, these regulations are generally regarded by the EU as great successes (Mazzini and Scalzo, 2023). However, this view is not shared by all experts. The use of technical standards and the CE mark, even outside AI, has already attracted its fair share of criticism in the literature, as there is no guarantee that a product bearing the CE mark will actually be "safe".

This was clearly demonstrated during the PIP scandal, named after the Poly Implant Prothèse company in France. The case came to light in 2010 after the PIP company, known for providing silicone breast implants, was accused of failing to comply with quality standards in the manufacture of some of its implants. At the time, only one type of silicone gel was allowed for breast implants and PIP, in order to cut costs, manufactured some of their implants with a mixture of this gel and a sub-standard industrial silicone gel (van Leeuwen, 2014). The French public control agency received signals that some of the PIP implants were causing health issues to the women bearing them, with an alleged risk of breast cancer, leading to their withdrawal from the market and the liquidation of the company. The implants were certified and CE marked by TÜV Rheinland. Indeed, CE marking for class III medical devices, which include breast implants, requires a third party certification by a notified body (Rott, 2019). The corresponding EU Directive also imposed surveillance duties on the notified body, such as periodically carrying out appropriate inspections to the manufacturer's quality system, and TÜV Rheinland is alleged to have failed[75].

The Pendra case is another example of damage caused by a CE marked product. It concerns a glucose monitoring device that was prematurely marketed by the Pendra company in the Netherlands in 2003. The device was not suitable for a large number of people due to the different properties of the skin and underlying tissues (Wentholt et al., 2005). Although details of the evaluation process have not been disclosed[76], some believe that the notified body should have identified this alleged defect.

While products subjected to external controls can be still be defective, this is even more true for products where no third-party audit is required before affixing the CE mark. As such, toy recalls in the EU, which uses CE marking and self-assessment, are ten to twenty times higher than in the US, where toy certification requires independent third-party certification (Larson and Jordan, 2019). As most high-risk AI systems will fall under this self-assessment procedure, some fear that this will not be enough to guarantee the safety of AI systems (Wachter, 2024).

[75] First, national courts of Germany and France reached different conclusions about the scope of the obligations of certification bodies and their possible liability in case of damage and harm (van Leeuwen, 2014). The extent of the notified body's duties was finally addressed in the CJEU case of Elisabeth Schmitt, where the CJEU confirmed notably that the notified body is for instance *"not under a general obligation to carry out unannounced inspections"*.

[76] This was seen by some as a lack of transparency that directly affects the consumer's right to information (Wentholt et al., 2005).

3.6.5 The difficulties in effectively implementing obligations

Finally, some of the criticisms also relate to the implementation of the regulation in practice. Some are particularly concerned about the lack of institutional strength for effective enforcement[77], or the lack of democratic supervision. Given that an AI system will pass through many hands between the time it is produced and the time it is actually deployed, some also worry about the allocation of responsibilities along the value chain (Edwards, 2022). In addition, the instability of the technology could create a *"pacing problem"* (Marchant, 2011) whereby systems developed after regulation's entry into force will not be properly covered. This was witnessed with the provisions on GPAI, which were added at the last minute, after GenAI boom the year before (Almada and Petit, 2023). Yet, there is no guarantee that such situation will not happen again in the future.

Furthermore, critics worry about the lack of complaint mechanism and the fact that the AI Act does not facilitate recourse (Ebers et al., 2021). However, this issue is partially addressed in the final version of the AI Act, with Article 85 allowing individuals or groups of individuals to lodge a complaint with a market surveillance authority and Article 86 providing for a right to explanation of individual decision-making. Despite the limited scope of these two rights, scholars consider this to be a positive development in the AI Act, moving towards a right to explanation that was virtually non-existent beforehand (Wachter, 2024).

Criticisms are generally addressed to the regulation itself as well as the means of compliance, but there are also concerns over the expected impact for industry and civil society (Vainionpää et al., 2023). Indeed, although a vast majority of the literature calls for more stringent requirements, another part worries about the potential compliance cost. This cost could range from a few thousand euros for the compliance of one AI system, to several hundred thousand euros for setting up a quality management system, as requested by Article 17 (Haataja and Bryson, 2021). However, this cost will only apply to large companies, as micro-entreprises may comply with certain elements of the quality management system in *"a simplify manner"* (Art. 17). Nevertheless, SMEs could still suffer from this compliance cost (Hacker, 2024).

3.7 A distinctive approach to measuring risks and ensuring rights

Despite other criticisms, the aspect of the AI Act that has really been at the center of debate is that of framing the text not according to the rights of individuals, but on a compliance framework that examine a level of risk to fundamental rights. In this section, we analyse what risk management looks like in the AI Act and why conflating risks and rights could weaken the protection of fundamental rights.

3.7.1 Risks in the AI Act must be reduced to an "acceptable" level

Following policy documents recommendations and a series of political choices, the AI Act has adopted a risk-based approach to the categorisation of systems. This risk-based structure implies that some applications of AI systems are "acceptable", while others – the "unacceptable risk systems" – are not (Laux et al., 2023).

But this risk-based approach, and its sister notion of acceptability, is not limited to the systems category; it can also be found directly in the requirements of the legal text, particularly in Article 9 on risk management. The risk management re-

[77] This was reinforced in the final version of the text, with the creation of the AI Office, the AI Board and the whole governance ecosystem. For more information, see Section 3.5.4.

quirement is central in the AI Act (Schuett, 2023b). It is the first requirement set out for high-risk systems, and arguably encompasses all the other requirements. According to Article 9, providers of high-risk AI systems must establish, implement, document and maintain a risk management system, comprising notably of the identification of known and foreseeable risks and the adoption of appropriate measures to eliminate or mitigate those risks. Following mitigation measures, residual risks must be reduced to an "acceptable" level.

Fraser and Bello y Villarino (2023) examine what "acceptable" risk might mean in the context of the AI Act. In their view, the European Commission encourages risk reduction *"as far as possible"*, i.e. insofar as it is feasible, whatever the costs. The European Parliament's amendments (European Parliament, 2023a) tended instead to encourage risk reduction *"as far as reasonably possible"*, weighing up costs and benefits before deciding on a threshold of acceptability. The final version of the AI Act (European Parliament and Council, 2024b) balances these two versions as risks must be eliminated *"as far as technically feasible through adequate design and development of the high-risk AI system"* (Art. 9(5)(a)). The acceptable level of risk will therefore be dictated mainly by the state of the art.

3.7.2 *The AI Act between risk- and rights-based approaches*

This structure around risk and acceptability determined by the state of the art is not new, since the GDPR, for example, followed the same pattern. This time, however, risks are considered in the AI Act as regards to the *"heath, safety and fundamental rights"*. The text also shows the intention of the European institutions to put in place various tools – standards, codes of practice, impact assessments and so on, to identify and prevent potential violations of fundamental rights. But in doing so, manufacturers will have to determine how to measure a level of risk to fundamental rights, as well as a level of acceptability of such a risk. The question is therefore whether it is possible to reconcile a risk-based approach with a more traditional right-based approach.

Risk-based regulatory approaches take their roots in the safety of critical infrastructures, but they have recently been widely applied in the context of digital technology regulation, such as the GDPR, DSA or the AI Act (Maxwell, 2022). They follow the philosophical movement of "utilitarianism", trying to maximise benefits by balancing economic interests and social well-being through quantitative analysis. This calculation takes the form of risk assessments, which help support the regulation, by providing a means to identify, assess and control risk. Risk is then understood as the combination of the probability of occurrence of an harm and its severity[78] Risk-based regulation is generally seen as a flexible and *"functionally efficient"* tool, designed to accompany a culture of risk management within companies (Black, 2010). Rights-based approaches, on the other hand, are not clearly defined in the academic literature but can be considered to be based on fundamental rights and the philosophical movement of "deontology", and place the individual at the center of moral and legal debate[79]. In rights-based regulation, rights are considered non-negotiable and must be respected regardless of the level of risk (Hidvegi, 2021). As a result, rights-based approaches generally consider that a violation of fundamental rights is not quantifiable, that all violations are reprehensible and that there is no trade-off with economic benefits whereby a risk to these rights would be acceptable (Maxwell, 2022).

[78] This is the definition usually presented in European law such as the AI Act (European Parliament and Council, 2024b, Art. 3(2)) or the Product Safety Regulation (European Parliament and Council, 2023c), but it is also a widely accepted definition in risk management (Aven, 2016), where the focus is more on business risk.

[79] See in particular the work of Rawls (1971).

However, a rights-based approach to AI is complicated to put in place as it requires effective enforcement mechanisms (Smuha, 2021a). Furthermore, different AI technologies may present different issues and should be regulated differently, which is why many scholars have also emphasised the need to regulate AI through risks (Schuett, 2023a). Others believe that AI regulation should be built on the pillars of liberal democratic societies: fundamental rights, the rule of law and democracy (Smuha et al., 2021). Similarly, Ruschemeier (2023) argues that the regulatory efforts for AI should be focused on enhancing the protection of legal rights, in particular by enacting the precautionary principle. In response, the AI Act attempts to accommodate both, mixing the semantics of the rights-based approach with the mechanisms of the risk-based approach. The place of fundamental rights in the AI Act is therefore still being debated (Almada and Petit, 2023), with some believing that it is no more than a marketing tool, and that economic benefits with the improvement of the European internal market are the predominant objective, relegating fundamental rights to second place (Castets-Renard and Besse, 2022).

3.7.3 *The risks of measuring risks*

Risk management is based on two assumptions: first, that it is possible to anticipate every risk; second, that for each risk, it is possible to calculate an associated probability and precise magnitude.

However, neither of these two assumptions is true in all cases. The first one because there is no such thing as zero risk. People assessing the risks might miss situations in which a risk could occur. This is particularly true of risks that are specific to a certain minority group. The fewer people affected, the less likely the risk is to be identified. These frameworks therefore render invisible certain kinds of harms suffered by certain groups of individuals (Kaminski, 2022). Risk measurement is inherently unfair today, as there is often insufficient data on certain groups – such as women or ethnic minorities – to properly assess risks. Drug doses or seatbelt strength are measured on a "general" body type that is not representative of every individual (Perez, 2020), and will only work on the statistically largest or most powerful group. The second assumption is based on a measurement paradigm: the idea that any observable phenomenon can be evaluated in a quantifiable way. But risk measures are often approximate and, by hiding behind "scientific facts", can give a false sense of accuracy (Rothstein et al., 2006).

Risk management also implies that the technology will be adopted despite its harm (Kaminski, 2022). As such, risk assessments can sometimes be used to justify a policy decision that was already made rather than truly giving insight on what should truly be done (Rothstein et al., 2006).

These general conclusions about risk frameworks have even greater implications when the frameworks are applied to fundamental rights. Indeed, both the identification and the measurement of the risk to fundamental rights are highly biased depending on who carries out the assessment. Risk measurement, in particular, can only be carried out using proxies, such as algorithmic unfairness for discrimination, and therefore misses out on a large proportion of possible risks.

3.7.4 *Defining a threshold of acceptability for fundamental rights: an impossible task?*

Defining a threshold of acceptability is a difficult issue in risk management. For instance, in the case of the safety of a nuclear power plant, the risk of an accident are quantified, the benefits of nuclear energy and the costs are weighed up, and the plant is allowed to open if the measured risk is below a certain threshold: the "acceptable level of risk" to safety (Fischhoff, 1983). This threshold cannot be zero, because a "zero risk" approach would lead to the total rejection of nuclear activity, which could otherwise provide benefits to society. Compromise are thus necessary. While there is no direct threshold for the safety level of a nuclear power plant, this safety level itself depends on the threshold granted to the various safety components. To take another example, in the case of a car, standardised seat-belt robustness tests are carried out to assess that the risk of dying in a car accident is less than a certain probability.

However, even if such a safety threshold is commonly used in product safety regulations, its adaptation to fundamental rights, such as a discrimination threshold, is not straightforward. For instance, if fairness allow us to measure some notion of discrimination, then residual biases or unfairness must be understood as the expression of residual risks as per the AI Act (Orwat et al., 2024). Setting an acceptable level of risk of discrimination may then involve defining a threshold for fairness metrics. There are, however, many different measures of fairness that can be used in different situations (Barocas et al., 2021), some of them are sometimes even incompatibles (Chouldechova, 2017a), and it is impossible to define an universal choice. Indeed, fairness is highly context-dependent (Wachter et al., 2021b) and is therefore hard to standardise (Bringas Colmenarejo et al., 2022)[80].

Additionally to these difficulties, setting a threshold for fundamental rights also poses a question of where to draw the line. It is usually admitted that fundamental rights follow logic of optimisation (Alexy, 2010). This means that they must be extended to a maximum and only the least restrictive solution on fundamental rights should be accepted. However, in the AI Act, there is a shift from this "optimising logic" to a "satisfactory logic" whereby any solution above a certain threshold could be deemed acceptable (Almada and Petit, 2023). These two types of logic – optimisation and satisfaction – translates into what Busch (2011) calls "olympics thresholds", i.e. the best possible, and "filter thresholds", i.e. better than a certain limit. If the satisfactory logic may be better in a context of technology uncertainty where the state of the art is dynamic and the "best" solution changes quickly (Almada and Petit, 2023), the optimising logic may better protect fundamental rights.

However, both of these rationales do not account for the fact that setting a threshold can be harmful in itself. Yet, scholars have shown that *"threshold theory"*, i.e. the science of associating a quantifiable level with the acceptability of a harm, is in fact a strategy of assimilation whereby science is used to justify damage (Liboiron, 2021)[81]. According to this view, which is closer to that of the advocates of right-based regulation, rights and wrongs should not be quantified. Nevertheless, this is not the approach taken by the Commission with the AI Act, which assumes that thresholds need to be set for the risks of AI.

[80] However, some initiatives to standardise fairness still exist. We review them in Chapter 6.

[81] Liboiron (2021) associates threshold theory to colonialism, as thresholds are used to justify the pollution of indigenous lands.

3.7.5 Reconciling rights and risks: a distinctive approach which challenges the operation of the NLF

The distinctive approach to AI regulation led by the EU with its AI Act, mixes risk-based and rights-based regulation and proposes to quantify the unquantifiable, i.e. to set a level of "acceptable" risk to fundamental rights. This dual approach, which was also noted by a number of scholars (Ho-Dac, 2023; Almada and Petit, 2023; Gornet and Maxwell, 2024; Smuha and Yeung, 2024), is due to the fact that the AI Act is product safety regulation where compliance with the state of the art – for instance technical standards – is supposed to ensure a level of protection of the product consumer, here, the end-user of an AI system. Yet, while these requirements are generally considered to relate to safety or health, the AI Act also aims to protect fundamental rights. The objective of the AI Act –to protect fundamental rights –and the means it implements to achieve it –risk management and product compliance –are therefore not necessarily aligned (Smuha and Yeung, 2024).

For Almada and Petit (2023), the AI Act's approach to regulate through product safety and technical standards necessarily involves a "logic of evaluation", where risks should be kept below a certain threshold. On the contrary, the traditional approach to fundamental rights protection involves a "logic of proportionality", where risks should be minimised as far as possible. Ebers et al. (2021) further notes that the inclusion of individual rights into the AI Act, especially those ex-post such as recourse mechanism, might challenge the approach of the NLF which is fundamentally ex-ante. Standards could be forced to define thresholds, either by setting them at the best known performance at the time the standard was drafted, which risks becoming obsolete very quickly, or at a certain level of acceptability which will necessarily be very arbitrary. Another solution could be to leave room for interpretation for judges to decide what "acceptable" means in a given situation. However, this would require the company to decide beforehand what "acceptable" means for them, with the risk that the two visions might not align.

3.8 Conclusion

In this chapter, we have presented the core of the European approach to AI regulation: the AI Act. This complex piece of legislation takes its roots in ethics charters and policy documents, following their advice to adopt a risk-based regulation with a special focus on "trustworthiness", i.e. respecting ethical and legal frameworks while being technically robust.

The risk-based approach was notably tested in the previous major success of European regulation in the digital sector: the GDPR. However, the AI Act goes further by proposing to adopt a product safety approach, whereby different risk categories of systems will have to comply with different legal obligations that will be assessed before the system is put on the market. These categories are defined in particular according to the sector in which the AI system is deployed, and the AI Act establishes a pyramid of these risks, ranging from higher risk with strict requirements, to lower risk with softer requirements. However, this classification is challenged by a cross-sectoral category: GPAI models, and its systematic risk subdivision, which include most of the current GenAI models. To ensure that obligations are met and that the AI Act is enforced, a whole ecosystem was designed to implement the various parts of the text, to enable the evaluation of systems and the deployment of measures in favour of innovation.

If the AI Act faced numerous criticisms when it first came out in April 2021, it was above all its dual approach, mixing risks and rights, that attracted attention the most. Indeed, the AI Act will require high-risk AI systems providers to put in place a risk management system and to reduce risks to a "technically acceptable level", which will depend on the state of the art, in particular the content of technical standards. This risk-based approach for fundamental rights will, however, inevitably pose problems when it comes to defining technical requirements in standards. It remains to be seen how far standards will go in making normative choices such as setting a threshold of acceptability. But even if they do not go that far, the mismatch between a risk-based approach, where compliance is assessed *ex ante* and a rights-based approach, where violations of rights are assessed *ex post*, could challenge the operation of the NLF.

Chapter 4
Technical standards: a product safety approach adapted to AI

Where does this chapter come from?

This chapter serves both as a background study on technical standards and product safety in Europe, and as an overview of how the AI Act fits into this context.

Sections 4.2 to 4.5 are an original contribution of this thesis. They give an introduction to technical standards and their role in the EU. Section 4.7 extends on the first section[1] of the following article:

> Gornet, M. and Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, 13(3). `https://doi.org/10.14763/2024.3.1784`

This section has been updated to take account of the current version of the IA Act and has been expanded to cover in greater depth the parts of the text relating to harmonised standards and CE marking. The remaining of the article is used for Chapter 7 and Chapter 5.

The reason for this separation is to ensure that we develop all three aspects of the original article: (1) how the AI Act deals with hENs and CE marking, examined in this chapter; (2) how the legal status of harmonised standards impacts the legitimacy of the ESOs, examined in Chapter 7; and (3) the AI Act's distinctive approach, compared to other product safety regulations, to address fundamental rights issues and what this means for technical standards, examined in Chapter 5.

Chapter's abstract

The use of standards as a regulatory mechanism has become a central means of ensuring product safety in Europe. In a broad sense, standards are technical documents defining voluntary guidelines for the industry, but in European law, they are the work of a few "experts" who come together in "recognised standardisation organisations" to converge on technical requirements for products. In this chapter, we look broadly at the field of standardisation, recounting its story, from the first global initiatives on the standardisation of electricity in the 19th century to the development of a European legal framework for standardisation in recent years. We show how the AI Act fits into this framework, relying heavily on standards for conformity assessment and CE marking, which is required for high-risk AI systems to be placed on the European market. The main role of standards has always been to capture industry expertise. This culminates in the AI Act, as the text further leaves AI providers the discretion to assess the conformity of their systems.

[1] The Section initially named "Protecting fundamental rights through product safety tools".

4.1 *Introduction*

We live in a standardised world. From our transports, accommodations, means of communication, to what we eat and consume, everything in our daily lives meets a certain number of standardised criteria that enable the operability of services and protect us as consumers. More specifically, when talking about standards, what usually comes to mind is the work of recognised organisations, such as the International Organization for Standardization (ISO), which are dedicated to producing technical documents setting out these standardised criteria. Furthermore, in Europe, the product safety framework has used these technical documents since the New Approach in the 1980s, a system that has only been strengthened in recent years with the New Legislative Framework (NLF). Products covered by the NLF must comply with a certain number of safety requirements. But while the law only sets out general "essential requirements", technical standards, and in particular the harmonised standards requested by the European Commission, will be responsible for specifying these requirements using more precise technical criteria. In summer 2024, European institutions published a new legal text for the regulation of Artificial Intelligence – the AI Act (European Commission, 2021e) – which makes use of this standardisation system for the conformity assessment procedure and the deliverance of the CE mark, which is necessary for certain types of AI systems, considered "high-risk", to be distributed on the European market. But what exactly are these standards on which the AI Act relies and how are they used in this new regulation for AI?

In this chapter, we give all the necessary background elements to understand the relationship between standards and European law, and specifically the AI Act. We first explain, in Section 4.2 to 4.4, what are technical standards, what purpose they serve, how they have evolved through time and how they are studied in the academic literature. We then present in Section 4.5 the ecosystem of standardisation, with the different organisations involved and the link with certification. In Section 4.6, we present the framework for product safety in the EU and how it makes use of technical tools such as standards and CE marking. Finally, we lay down in Section 4.7, how the AI Act makes use of the product safety regulatory approach, including how standards are used for conformity assessment.

4.2 *What are technical standards*

4.2.1 *Defining standards*

There are many ways of defining what a standard is, because the word "standard" is polysemous and can have different meanings in different contexts. In the general sense, a standard means *"something established by authority, custom, or general consent as a model or example"* (Merriam-Webster Dictionary, n.d.j). It is a criterion which can serve as a benchmark for comparison. Standards can be codified, for example in codes of conduct or professional standards, or they can be assessed based on experience. For example, performance standards in computer science are not described in one specific document but can be deduced from the general state of the art made up of a multitude of scientific articles. As such, there is no telling what is exactly the level of such a standard, it is a flexible reference point. Standards, understood in this general sense, are often seen as the usual way of doing something, but also the best way. But in reality, following standards is first and foremost the most accepted way of doing something (Brunsson and Jacobsson, 2002). For ease of understanding, we will refer to this general meaning of standards as "common norms"[2].

Indeed, in this thesis, we are specifically interested in the second sense of the term standard: *"a rule for the measure of quantity, weight, extent, value, or quality"* (Merriam-Webster Dictionary, n.d.j). This is when a standard ceases to be a vague shared notion of what can or should be achieved and begins to be codified in technical terms. We will refer to these codified documents as "technical standards", or simply as "standards"[3].

The term "standard" is defined differently in academic literature and in documents distributed by standards bodies themselves or regulatory entities. For economic scholars, standards are *"documented agreements containing technical guidelines"* (Allen and Sriram, 2000). On their website, the International Standardization Organization (ISO) refers to them simply as *"a formula that describes the best way of doing something"* (ISO, n.d.h). Technical standards notably serve to establish *"common solutions"* that every company in the same field can use (Cellucci, 2008) and ensure that products are *"fit for their purpose"* (Allen and Sriram, 2000).

Some definitions go further by defining what standards can contain. For instance, the Circular n°A-119 of the Office of Management and Budget (OMB) (OMB, 1998) defines them as a *"common and repeated use of rules, conditions, guidelines or characteristics for products or related processes and production methods, and related management systems practices"*, which can include *"the definition of terms; classification of components; [...] specification of [...] performance, designs, or operations; measurement of quality [...] ; test methods [...]"* and so on. Similarly for ISO, standards cover a wide range of activities, like *"making a product, managing a process, delivering a service or supplying materials"* (ISO, n.d.h).

Most of the experts in standardisation groups come from the industry[4], therefore they know the field and they know the market needs. As ISO puts it, their standards are *"the distilled wisdom of people with expertise in their subject matter and who know the needs of the organisations they represent"* (ISO, n.d.h). The selling point of standards is therefore the expertise of the people who develop them.

But ISO has not contented itself with simply explaining what a standard is on its website, it has – quite obviously – standardised the definition itself. The first definition of "standard" in a standard appears in the ISO/IEC Guide 2, in 1976 and was later revised in following editions[5]. ISO/IEC Guide 2 defines a standard as *"a document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of die optimum degree of order in a given context"*[6] (Brunsson and Jacobsson, 2002). Brunsson and Jacobsson (2002) further ironises that *"fortunately, since this definition is a standard, its acceptance is voluntary"*. But sarcasm aside, there are real reasons to consider other definitions. Indeed, the most important objection is that this definition limits the use of the term "standard" to the products of "recognised bodies"[7], i.e. recognised Standards Developing Organisation (SDOs).

For this thesis, we choose to broaden the scope of this definition[8]. Therefore, we will use our own definition of the term "standard", adapted from (Frattone, 2022):

[2] Not to be confused with legal norms, which are binding rules.

[3] Standards usually tend to reflect common norms and practices, but this is not always true (Brunsson and Jacobsson, 2002). Innovation and market dynamics also play an important role in standard making. For more information, see Section 4.3.

[4] For ISO participation, see (Morikawa and Morrison, 2004). For participation in CEN-CENELEC JTC 21 on AI specifically, see our data analysis in Chapter 8, Section 8.4.1.

[5] The first to appear on ISO's website is the seventh edition of Guide 2, published in 1996. The Guide was then was revised in 2004 and has also been adopted in Europe under the name EN 45020:1998, then EN 45020:2007. However, we do not have access to this standard, therefore we found the definition in (Brunsson and Jacobsson, 2002) who directly quotes EN 45020:1998. This definition of standard should be still valid as ISO states in the "foreword" to the eighth edition (2004), that it is identical to the seventh edition (1996) except for the definitions on conformity assessment.

[6] A similar definition is also given in the WTO Technical Barriers to Trade agreement (WTO, 1995).

[7] We have been unable to find any list or definition of what a "recognised" organisation is for ISO, but it is possible that they appear in the rest of Guide 2 which we do not have access to. However, a list of SDOs "recognised" by Europe is provided by Regulation 1025/2012 (European Parliament and Council, 2012). For more information, see Section 4.6.2.

[8] See the full justification in the box in Section 4.6.2.

> ***Definition***
>
> ***Standard.*** *A document containing technical requirements or guidance, addressed to professionals, which codifies industrial expertise and compliance with which is voluntary.*

4.2.2 Classifying types of standards

Researchers have come up with various classifications for standards. One of the most widespread is the binary division between *de jure* and *de facto standards* (David and Greenstein, 1990), also referred to in some works as *formal* and *informal standards* (Farrell and Saloner, 1985). However, this classification differs from author to author. For instance, Allen and Sriram (2000) prefers to separate the usual de jure category into *regulatory* and *consensus standards* according to whether they are created by regulatory agencies or voluntary standards bodies. The same type of classification in three categories rather than two is made by Kaplinsky (2010) who distinguishes between *private sector*, *state sector* and *civil society sector standards*.

Whilst de jure standards emerge through industry consensus, de facto standards are dictated by market law (Stango, 2004). De jure are obtain after negotiations between stakeholders whilst de facto is the last standing option when one company has imposed its solution over the global market. For instance, standards developed in the EU by European Standardisation Organisations (ESOs) are automatically de jure standards[9] whereas the QWERTY standards for keyboards is an example of a de facto standard that was initially developed by a private company and that has established itself in the market (Allen and Sriram, 2000). It should be noted however that de jure standards that are voluntary under EU law for instance, can become de facto rules for the marketability of products (Hofmann, 2016). Even if not mandatory at first, if these standards are widely adopted by companies, the ones that fall behind can be shunned by potential clients.

Another distinction that can be made is on the content of these standards. Several works agree on a common classification between broadly four types of standards: *information*, *variety reduction*, *minimum quality and safety*, *compatibility and interoperability* standards (Blind, 2004; Swann, 2000; Tassey, 2000). This exact terminology comes from Blind (2004), Swann (2000) and Tassey (2000) use slightly different terms but the categories stay the same. However, the classification stems from David (1987) who originally distinguished between standards according to the economic problems they solve: *"for reference and definition"*, *"for minimal admissible attributes"* and *"for interface compatibility"*. Information standards contain terminology as well as *"test and measurement methods for describing, quantifying, and evaluating product attributes"* (Tassey, 2000). As a result, they help stakeholders speak a common language and define best practices. They can also contain state-of-the-art methods or reference certain equipment used either in the design or testing phase. Variety reduction standards limit a product to certain characteristics, often with the desire to reduce the types of products on the market. Minimum quality standards, also called *performance-based standards* (Allen and Sriram, 2000) aim at specifying a threshold under which a technology should not be used for safety reasons. It is used as a starting point for companies who then compete on higher level of performance or for this base level of performance lower the price to a minimum (Tassey, 2000). Finally, interoperability standards specify properties that products need to meet to work with larger systems.

[9] See Section 4.5.1 for more information about the ESOs.

Standards can be classified as either horizontal or vertical. Horizontal standards deal with issues that apply across several areas of technology whilst vertical standards address sector-specific issues (Nativi and De Nigris, 2021). As such, broad information standards defining terms and concepts are usually horizontal standards whilst performance standards are usually focusing on a specific technology.

Lastly, some scholars also distinguish standards on the object of standardisation: usually *product* and *non-product standards* (Tassey, 2000). Non-product standards can also be referred to as *process standards* (Kaplinsky, 2010). This binary classification is strongly linked to the type of requirements present in standards. For instance, in the context of AI, Laux et al. (2023) distinguishes between *procedural* requirements, such as transparency, which facilitate stakeholder accountability but are carried out by non-experts, and *substantive* requirements, such as robustness, which require technical expertise. Procedural requirements are therefore defined mainly for processes and substansive requirements for products. Process standards are also sometimes called *management* standards (Yates and Murphy, 2019)[10].

4.3 *What is the role of technical standards*

In the past decades, the role of standards has strongly increased, with ever growing economic rewards, more demands for the quality of products and interoperability of systems Tassey (2000). But these standards can serve different purposes.

4.3.1 *Standards and interoperability*

For the average person, standardisation has been a good thing, participating in making the world work just a little better (Yates and Murphy, 2019). Yates and Murphy (2019) begin their book by explaining that the pages the reader holds in their hands are a standardised paper format and that, if they are reading a digital copy, the e-reader software, the screen, the device's battery and the power plant which gave it electricity, are likely to be standardised. Standards encourage products convergence and consumers are less likely to be confused by the variety of goods on offer. Standards are therefore said to make *"a complex world simpler"*, although this is usually a marketing argument made firstly by standardisation organisations themselves (Brunsson and Jacobsson, 2002).

But sometimes, product convergence and agreements between the various stakeholders on the solution to be adopted are necessary for complex systems – requiring interaction between several products – to function properly. For instance, a phone is more useful when it can communicate with any other one, regardless of the manufacturer **?**. Similarly, printers must have the same input paper format. And even if there is no perfect convergence between electricity formats around the world, it has almost become necessary in the same region to have the same plug format to plug in electronic devices without always needing a converter. Interoperability has therefore always been the primary and most important objective of standardisation.

Finally, this convergence of solutions is also useful for simplifying development processes. Indeed, standards help companies to adapt their product to the constraints of the environment of use. The benefits of this adaptation is to promote cost optimisation by increasing efficiency (de Vries, 1999). Once there is consensus on a standard, every process follows the same methodology, all systems are built and tested in the same way and productivity is improved. Products are less expen-

[10] Either quality management standards, or risk management standards.

sive for companies to manufacture and certifiers can stick to simpler procedures. However, adapting an entire company to a new standard is time consuming and expensive. The cost efficiency can thus only be visible on the long run, requiring companies to invest at first (Graz, 2006).

4.3.2 *Standards and market dynamics*

Standards can also help boost the economy. As early as the 2000s, the German Institute for Standardisation (DIN), published a study, based on a macroeconomic analysis, showing that the economic benefits of standardisation was about 1% of Germany's gross national product at the time (for Standardization e. V., 2000). If this study has since then been revised downwards (Blind et al., 2012), the benefits for businesses and economy as a whole remain significant. This study as been reproduce in other countries around the world with similar conclusions (DTI, 2005; Haimowitz and Warren, 2007; Australia, 2006; Miotti, 2009).

However, if it is commonly admitted that standards have a positive effect on the economy, the same cannot be said of innovation. The literature is split between the idea that standards could inhibit innovation by codifying inefficient technology that can quickly become obsolete and the thought that it could also spur it by forming a baseline from which new technologies can emerge (Allen and Sriram, 2000). As such, the effect of standards on innovation depends on the context and content of these standards. Existing standards can put pressure on other companies to innovate (Blind, 2016). There may be a contest between companies to excel in a specific standardised test, as is the case in the field of facial recognition with the National Institute of Standards and Technology (NIST) benchmark (NIST, n.d.b) where companies compete to get the higher score. On the other hand, standards can also nurture cooperation between companies in the case of interoperability standards (Cihon et al., 2021).

Yet, if standards come too late, there is a risk of codifying a technology that is already obsolete and setting a yardstick that is too low in relation to the current state of the art. On the contrary, if standards come too early they can lead to premature and inefficient selection of technology (Blind, 2016). This was the case of the QWERTY keyboard that despite its many drawbacks, has established itself in the market at the expense of innovation (Allen and Sriram, 2000). This might happen for instance when the market is dominated by a single company.

If standards affect the market dynamics, the inverse is also true: market dynamics can determine which technology will become the dominant version and turn into the next de facto standard (Tassey, 2000). This encourages competition amongst companies who each want to impose their version of the product. With digital tools, this phenomenon has taken on a new dimension: in a market where the GAFAM[11] prevail, enforcing a different standard than theirs has become a challenge and a matter of national governance for companies outside the US.

4.3.3 *Standards and consumer protection*

In general, standards help companies demonstrate that the product indeed possess the innovative features they claim to have or that the level of risk in behind a certain threshold (Swann, 2000). Safety standards thus provide assurances that the consumer is not likely to be injured or killed when they use a product (Hanson,

[11] Google, Amazon, Facebook, Apple, Microsoft: the big American technology companies.

2005), such as the crashworthiness standards for cars. Additionally, by reducing these risks, standards can contribute to build trust in the product (Blind, 2016). Reardon et al. (1999) describes standards as *"credence goods"*: the information they provide can help make consumption decisions by displaying *"quality and/or safety aspects that cannot be known to consumers through sensory inspection or observation-in-consumption"*. As such, quality or safety standards permit to create transparency in the market (de Vries, 1999) and enable the freedom of choice for users (Cantero Gamito, 2018). Following a standards can then be seen as a mark of quality or security. Standards also help codify and diffuse the state-of-the-art (Swann, 2000), bringing each new similar product to a minimum quality level and protecting the consumer from unreliable products. They can also serve as a mechanism for evaluating and comparing systems, pushing manufacturers to put in place safeguards against potential hazards and thus further protect the consumer.

However, when these standards are too weak, consumer protection can backfire. This is particularly the case with performance standards, which can serve both as a threshold that companies seek to achieve or as a minimum requirement that they already have and do not need to exceed, thus acting as a double-edged sword.

The EU is particularly committed to promoting fundamental rights and values, down to its legal assessment (Williams, 2009). For this reason, standards are seen as a means of fostering EU values and increasing consumer confidence (European Commission, 2022e). For instance, recently the European Parliament mandated the use of a common charger for mobile devices. The stated aim is to avoid negative impacts on consumers and the environment, by making it easier to see whether chargers and devices are compatible and reducing electronic waste (European Parliament, 2022). This approach is a way of reaffirming Europe's commitment to the fight for environmental and social rights.

Standards do not replace traditional means of compulsory regulation. However, they can help to complement legislation by providing a technical perspective that the law lacks and by being updated more quickly. Indeed, regulators lack the expertise to assess these technologies themselves and are usually subject to budgetary restrictions that prevent them from recruiting the best experts (Baram and Bieder, 2022). Furthermore, it is not the role of the law to incorporate technical elements which can quickly become obsolete. In addition, the development of new standards, although a lengthy process, is always faster than the drafting of new laws. They are also frequently revised[12] to incorporate new elements or to correct the state of the art in line with developments in the field. Standards therefore allow for more flexibility than the traditional legislative process. This makes them particularly suitable for the regulation of new and rapidly evolving technologies (Frattone, 2022).

4.4 *Standards in the academic literature*

4.4.1 *A short history of standardisation*

Standardisation has been around since humans have tried to make things work together. Some date standards to the bronze age with the standardisation of exchange values to facilitate trade (Oever and Milan, 2022). For some others, standardisation was most visible during the world wars, where each side applied its own standardised techniques between allies (Hesser and Czaya, 2010). However, at the time, these initiatives were not officially produced by recognised bodies. The

[12] ISO standards, for example, are updated at least every five years, and may be revised sooner if there is a major change in the field.

end of the 19th century was particularly rich in standardisation initiatives, as it was a time of rapid technological innovation. With the Second Industrial Revolution came standards for product safety, interoperability and performance. Objects of standardisation were as diverse as steam boilers, screw threads and steel rails. A wide range of organisations began experimenting with different mode of standard setting before converging around 1880 towards a what economists have called *"standardisation by committee"* (Yates and Murphy, 2019).

One of the first concerns at the time was how to standardise the measurement of electricity (Loya and Boli, 1999). European countries in particular organised several international congresses on the measurement of electrical resistance, as the lack of a standardised measurement was seen as the main obstacle to the wider use of electricity. The International Electrotechnical Commission (IEC) was therefore one of the first standardisation organisations to see the light of day in 1906. In non-electrical fields, the movement was more difficult to set up and several organisations were founded and collapsed, until the creation of the International Organisation for Standardisation (ISO) after the war in 1947. But ISO mainly forged its reputation in the 70s and 80s, by developing the first international standards following the expansion of international trade (Dumitru, 2008). One of its best-sellers at the time was ISO 9000, a standard for company management procedures and production, for which ISO provided certification. Certification to the ISO 9000 standard was said to boost economic performance and customer satisfaction. It was adopted by tens of thousands of companies worldwide, each one bringing a new success story that fuelled the reputation of the standardisation organisation (Loya and Boli, 1999).

The types of standards that where developed by standardisation organisations also evolved through time. Yates and Murphy (2019) distinguish three main waves of standardisation: the first centred on national initiatives, the second with the rise of international organisations such as ISO, and the third emerging with digital technologies and based mainly on consortia standards and a diversification of international standards. Indeed, professional organisations have long played a major role in standard setting, usually by defining their own codes of conduct to regulate a profession (Hutter, 2006). But more recent organisations, such as the Internet Engineering Task Force (IETF) or the World Wide Web Consortium (W3C), have gradually established themselves, particularly for new technologies, as an alternative to traditional standardisation organisations such as ISO (Yates and Murphy, 2019). Moreover, the more "traditional" standardisation organisations have also open their doors to various profiles. While in the late 19th century, all the people in standardisation were engineers or skilled executives, the community of standards setters today has enlarged, with people from very different background (Yates and Murphy, 2019). Standards are therefore the result of cooperation between governmental, private and civil society groups, which leads Graz (2019) to describe the organisations that produce them as *"hybrid authorities"* and Wiegmann et al. (2017) to speak about *"multi-mode standardisation processes"*.

In addition to the diversification of standards setters and processes, the third wave of standardisation has also seen a diversification in the content of the standards produced. After the success of ISO 9000 in the 1980s, organisations focused on developing standards based on this model, called *"quality management standards"* (Yates and Murphy, 2019). These standards address the processes and management practices implemented in companies rather than the technical characteristics of their products. Quality management standards have been declined for a variety of tasks and purposes, and have notably more recently been used for social and environmental standardisation[13] and in general for *"good governance"* (Steffek and Wegmann, 2021)[14]. But ISO was not the only one to shift from products standards to process standards, as this trend can be observed in many standardisation organisations at international and national level. An example of this trend is the National Institute of Standards and Technology (NIST) in the US, which has a long history of developing software standards, and has focused in recent decades on open-ended frameworks[15] (Choi, 2024). The standards of the modern age are therefore not concerned with products, but rather with the means of production. Standards relating to the working environment do not refer to the working environment itself, but to the procedures that companies must put in place to deal with its related issues (Brunsson and Jacobsson, 2002). Similarly, quality standards do not refer directly to the quality that is actually achieved, but rather to the processes that are supposed to lead to quality. Product requirement standards are therefore increasingly being abandoned in favour of the routine that a producer should employ to develop his product and determine for himself whether it is fit for market deployment.

4.4.2 *Early works on standardisation in economy*

Standardisation has long been a neglected area of research, specifically in social sciences (Brunsson and Jacobsson, 2002; Schepel, 2005). The first studies on standardisation were carried out in the field of economics. In particular, research on the economics of standards in information industries has been expanding rapidly during the 1980s (David and Greenstein, 1990). This literature focused almost exclusively on "compatibility" standards, what we would today call "interoperability standards", which were very prominent at the time, particularly in the telecommunications industry (David and Greenstein, 1990). In the 1990s, the literature continued to focus on compatibility, looking in particular at the different possible outcomes where there are economic gains from achieving coordination among stakeholders. Scholars have studied the plausibility of emergence of standards to ensure the compatibility of system components (David and Greenstein, 1990), the consequences of compatibility on technological progress (Matutes and Regibeau, 1996), and how companies could make use of such compatibility standards (Grindley, 1995).

Outside of interoperability standards, the majority of studies focused on the impact on standardisation on innovation (Allen and Sriram, 2000), or on the market as a whole (Tassey, 2000), and in particular on what economics have called the "standards war" (Shapiro and Varian, 1999; Stango, 2004), i.e. the competition between companies to impose their standard on the market. This has also lead to studies of technological development, such as that of the QWERTY keyboard (David, 1986), and whether the best standard always wins (Pinch and Bijker, 1984).

Finally, with the diversification in standards setting came, in the late 1990s, the first studies on standardisation organisations and on quality standards (de Vries, 1999), as well as the benefits of these standards for consumers (Reardon et al., 1999). In the early 2000s, some economic studies began to take an interest in the role of government in standardisation (Swann, 2000), paving the way for political scientists to take up the issue.

[13] Such as with ISO 14001 (2015a) on environmental management systems and ISO 26000 (2010) on social responsibility

[14] Although Steffek and Wegmann (2021) is not discussing a specific example, but rather how standardisation works in modern society, the most striking example of good governance standards is probably the recent ISO 37000 (2021) on the governance of organisations.

[15] An example of these open-ended frameworks is the NIST AI Risk Management Framework (NIST, 2023a).

4.4.3 *Standardisation as a means of governance and regulation*

The second field to take an interest in standardisation, in the late 1990s and early 2000s, was therefore political science. Researchers sought to analyse the power dynamics behind standardisation, questioning who sets the rules and who benefits from them. For instance, confronting standardisation with world theory, Loya and Boli (1999) analyses whether the biggest stakeholders always have the upper hand in standardisation. Schmidt and Werle (1998) have followed the negotiation processes of telecommunication standards and show how they are embedded in institutional structures. Others studies have analysed the externalities that lead to cooperation or competition on standards. For Abbott and Snidal (2001), standardisation can therefore be either a coordination problem, where actors are forced to find a common solution, or a Prisoners' Dilemma problem, where individually optimal behavior leads to collectively suboptimal outcomes. The category of problem depends, among other things, on the technological and regulatory conditions in which standardisation takes place.

Many authors frame standard setting as an important governance and regulation mechanism, far removed from the influence of the state (Abbott and Snidal, 2001; Brunsson and Jacobsson, 2002; Hutter, 2006; Graz, 2019). Abbott and Snidal (2009) speaks of a process of *"regulatory standard-setting"*, where collaboration between public and private sector is necessary. For Brunsson and Jacobsson (2002) standardisation organisations activities are a globalisation process in itself, where the increase of transnational activities and the growth of non-state actors leads to transnational companies taking over the role of the state. Standards are therefore perceived as instruments of control as they can be used to influence state policies.

With ISO's glory days also came a growing literature dedicated to analysing the work of international standards bodies and standards setting processes, as well as the role they play in world politics (Mattli and Buthe, 2003). Studies note a shift from local and domestic to international standard setting (Büthe and Mattli, 2010), and therefore a gap in the literature on the study of these institutional spaces. For instance, Büthe and Mattli (2010) draws up a typology of international standard bodies, depending on their public or private nature, as well as their market incentives, looks at the targeted audience and the participants in standard setting, and studies the standards adoption process of different organisations. Many of these works also try to draw attention to the role of private actors in standardisation (Mattli and Buthe, 2003; Brunsson and Jacobsson, 2002; Büthe and Mattli, 2011).

4.4.4 *Works on standardisation in European law*

While economists and political scientists have been studying standardisation since the beginning of the twentieth century, with an acceleration of academic work from the 1980s onwards, the legal field has been slow to take an interest. Even in Europe, studies on the European governance system based on co-regulation[16] were left to political scientists, who first analysed the framework of the New Approach that was beginning to take hold in the 1980s[17] (Pelkmans, 1987). Standardisation was seen as one mechanism among others in a new form of governance focused on removing the barriers to trade (Egan, 2001).

In legal studies, standardisation was first analysed from the perspective of com-

[16] For more information, see Chapter 1, Section 1.1.7.

[17] For more information on the New Approach, see Section 4.6.1 in this chapter.

petition law and intellectual property law, examining the compatibility of standards with patents (Staniszewski, 2007) and antitrust law (Lundqvist, 2014). A few studies, such as Bernstein and Hannah (2008), also examined standardisation from the perspective of international law, questioning the role of traditional international organisations and their trade agreements in regulating standards bodies. But standardisation can also be applied in private law and scholars have studied its impact in the European context (van Leeuwen, 2017). Indeed, standardisation lies on the borderline between public and private law (Gnes, 2017), and Schepel (2005) even believes that an approach exclusively from one of these sides is *"inherently incapable of capturing or explaining standardisation"*. In his work, Schepel (2005) conducted a large study looking at the place of standards in different countries across the world. In particular, he studied the European directives for internal market regulation and the role of standards which he perceived to be *"embedded"* in European law. In addition to this work, the European policy approach of "delegating" public rule making to private standardisation bodies has also been denounced by (Gestel and Micklitz, 2013). At the time, a number of important legal issues had not yet been addressed by case law, such as how far the delegation of power could go or who was responsible for the content of standards[18].

Despite efforts to highlight the importance of standards for European law in the 2000s and early 2010s, it took some time for public lawyers to take an interest in standardisation (Eliantonio and Cauffman, 2020). This changed in 2016 with the James Elliott ruling (CJEU, 2016a), in which the European Commission found that certain standards could be considered part of European law[19] and that the Court of Justice of the EU could interpret them[20]. Following this ruling, a plethora of studies analysing the *"juridification"* (Schapel, 2013) of standards emerged. While most of them start from the premises of the James Elliott case, numerous EU rulings since then have shown the special place of standards in European law[21].

4.5 The ecosystem of standardisation

4.5.1 The different organisations

SDOs are private entities whose business plan is to contribute to the development of standards so that they can then sell them[22]. However, despite their private nature, SDOs work closely with public bodies, as is the case in Europe[23].

There are six different official SDOs "recognised" by Europe[24]. Figure 4.1 presents this ecosystem and the interactions between SDOs. They are not the only entities to prepare standards[25] for AI[26]. However, as "recognised" SDOs, they are the key players whose work will support the AI Act.

To begin with, the best-known standards are often ISO standards, developed by the International Organisation for Standardisation (ISO). ISO was established in 1947 to promote international standards in all fields, except those covered by other organisations such as IEC. Nowadays, ISO standards are commonly used to certify the quality of the processes within a company[27], the company's environmental performance [28], or the security of its IT systems [29]. In the context of digital systems, ISO often collaborates with the International Electrotechnical Commission (IEC). IEC was established in 1906 to promote cooperation on electrical standardisation. Nowadays, ISO and IEC often collaborate on standard setting in Joint Technical Committees (JTC), such as ISO/IEC JTC 1 – Information Technology, which is host,

[18] If many of these questions are still unanswered, such as the copyright protection of standards, many others have be addressed in subsequent case law..

[19] These standards are called harmonised standards. For more information, see Section 4.6.2 of this chapter.

[20] For more information on the James Elliott case, see Chapter 7, Section 7.3.

[21] For more information, see Chapter 7, Section 7.3.

[22] These business incentives can pose certain problems, particularly with regard to access to standards and their copyright protection, especially for hENs that have a legal status in the EU. For more information, see Chapter 7, Section 7.3.

[23] For more information on the European context, see Section 4.6.

[24] See Section 4.6.2.

[25] According to our definition of the word "standard" in Section 4.2.1 which go beyond the work delivered by "recognised bodies".

[26] We give an overview of other initiatives in Chapter 5, Section 5.3.

[27] For instance ISO 9001 (2015b) on quality management systems and its associated standards (ISO, n.d.c).

[28] For instance ISO 14001 (2015a) on environmental management systems and its associated standards (ISO, n.d.b).

[29] For instance ISO/IEC 27001 (2022a) on information security management and its associated standards (ISO/IEC, n.d.a)

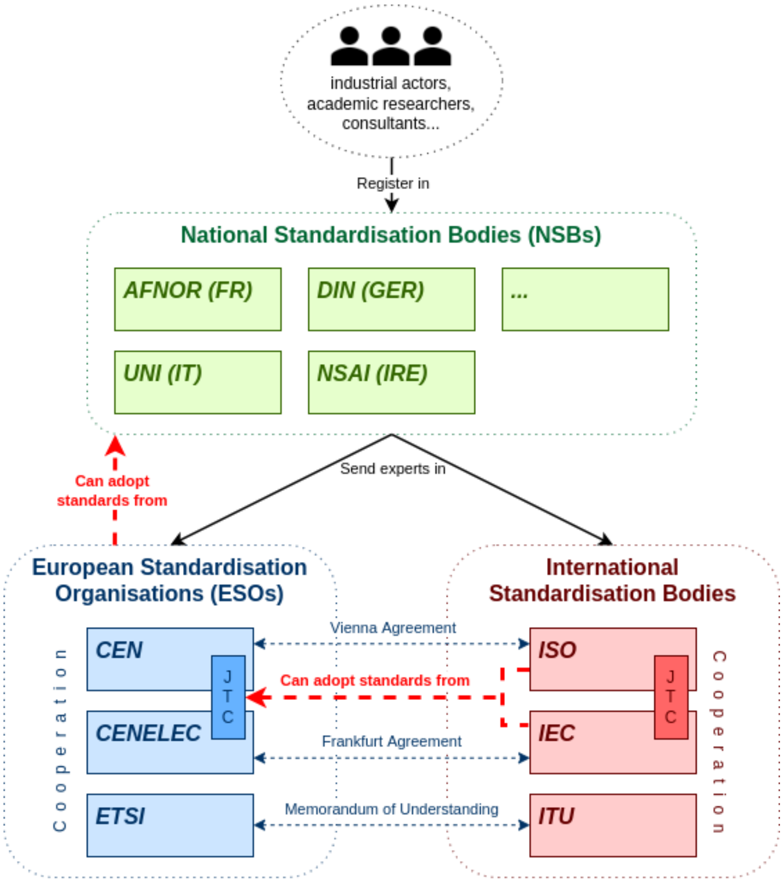


Figure 4.1: General ecosystem of standardisation.

among other things, to the work on AI [30] Finally, a third organisation, the International Telecommunication Union (ITU), is also involved in setting standards for information and communication technologies. The ITU was founded in 1865 by to promote the development of a telegraph network, and is now a specialised agency of the United Nations.

At European level, three bodies are responsible for developing standards: the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardisation (CENELEC) and the European Telecommunications Standards Institute (ETSI). These three bodies are known as the European Standards Organisations (ESOs). Founded in 1961, CEN is the European equivalent of ISO. It is active in all fields except those already covered by CENELEC. On the other hand, CENELEC was founded in 1973 following the extension of the single market to Denmark, Ireland and Great Britain (Dumitru, 2008), and is responsible for developing European standards in the field of electrical engineering. Finally, ETSI is the European equivalent of ITU, founded in 1988 and operating in the fields of telecommunication, data transmission, and IT in general (Dumitru, 2008). In the same way as ISO and IEC, CEN and CENELEC often collaborate within JTCs, for example on AI standards in JTC 21[31].

Additionally to "recognised" ESOs and international SDOs, at national level, each EU member state has at least one recognised standards body. These bodies are referred to as National Standardisation Bodies (NSBs) (European Commission, n.d.p)[32]. For instance, France has the *Association Française de Normalisation* (AFNOR), Germany has the *Deutsches Institut für Normung* (DIN)[33], the UK has the

[30] See Chapter 5, Section 5.3.2.

[31] For more information on the structure of JTC 21, see Chapter 5, Section 5.3.1.

[32] Note that this name is not reserved for European NSBs. Other countries around the world use the acronym NSB to designate their own recognised standards bodies.

[33] But also the *Deutsche Kommission Elektrotechnik Elektronik Informationstechnik* (DKE). To understand why there are two German NSB, see Section 4.5.2.

British Standards Institution (BSI), Italy has the *Ente Italiano di Normazione* (UNI), Ireland has the National Standards Authority of Ireland (NSAI), and so on. A list of all the NSBs recognised at European level is available at (European Parliament and Council, 2020).

European and international bodies are accustomed to collaborating on the development of standards. The six SDOs presented in Figure 4.1 have signed collaboration agreements to be able to attend each others meetings and collaborate on standards. The two most important agreements are the Vienna and Frankfurt agreements between CEN and ISO, and between CENELEC and IEC. These agreements enable the organisations to launch joint standardisation initiatives, as well as adopting each others' standards[34]. For example, International Standards (IS) originating from ISO/IEC can be endorsed by CEN-CENELEC and become European Standards (EN) following a vote by CEN-CENELEC members. Once adopted at European level, these standards can then be adopted, by ricochet, by national bodies. For example, AFNOR could adopt a CEN-CENELEC standard and it would become a French standard (FR)[35].

4.5.2 *Participating in standardisation work*

While some NSBs, such as the BSI in the UK, carry out standardisation work themselves, most work, for example on AI standards, is carried out at European and international level. As such, the main road to participate in standardisation work, both in international SDOs and in ESOs, is through NSBs. Standardisation experts register first in a NSB and are sent to work on standards at European and international level in various Technical Committees (TCs) and Working Groups (WGs). These experts are often industrial players, representing the interests of their parent companies. But experts can also come from research institutes or public establishments, for example. Therefore, anyone can apply to join a NSB to take part in standards development and standards voting, usually in exchange for a membership fee. These fees can, however, be waved on several occasions, such as for academics or SMEs, although it depends on the country[36]. The majority of the NSB's income therefore comes from the annual fees paid by standardisation experts and from the sale of standards. Additionally, for NSBs which also carry out certification activities, income may come from the sale of certification services.

Why use the term "standardisation experts"?
Several terms can be used to refer to the people who draw up standards. The academic literature, for instance, sometimes uses the term "standardiser" (Brunsson and Jacobsson, 2002), but this term is not officially recognised by SDOs. In this thesis, we choose the term "experts" because this is how the SDOs generally refer to the people who work for them (ISO, n.d.i). However, it should be noted that the field of expertise of these "experts" is rarely specified. Indeed, while some are technical experts in their field – e.g. computer scientists for IT standards – this is not always the case. Many of these "experts" have no technical background and are more experienced in the standardisation process itself. For more information, see Chapter 8, Section 8.4.1 and Section 8.6.2.

[34] However, such collaboration raises questions of sovereignty, particularly in the case of standards intended to support European law. For more information, see Chapter 7, Section 7.5.

[35] The letters preceding the name of a standard indicate the entity that endorsed it.

[36] In some countries, the law even requires any standards body to waive fees for SMEs, such as in France (Légifrance, 2009, Art. 14). In addition, some NSBs do not require *any* expert to pay a membership fee, for instance in Austria.

Each SDOs at European and international level has a list of NSBs "members" whose experts can take part in standardisation activities. However, each SDO has their own system for membership of NSBs, which differs from one body to another. In most cases, NSBs can be either full members with participation and voting rights, or associate members with a simple observer role[37,38].

For instance, in CEN or CENELEC, there are 34 countries that have a membership, included the twenty seven European Union countries, plus some selected countries[39]. Only one NSB can participate in CEN or CENELEC per country, except for the work conducted in JTCs where countries can have a NSB in CEN and one in CENELEC. This is the case, for example, for Germany[40]. But apart from members, other NSBs are invited to participate in CEN or CENELEC, with different levels of recognition: "Affiliates" are the NSBs of countries which are potential candidates for EU membership; and "Companion Standardisation Body" (CSB) are the NSBs which are also members of ISO or IEC. Affiliates and CSBs have free access to all CEN and CENELEC standards and can adopt them on national level, are observers in technical committees and can participate to assembly meetings although they do not have the right to vote. Additionally, "Partner Organisations" can participate in standardisation work at CEN and CENELEC. They are for instance European trade unions[41], consumer representatives[42] or professionals associations.

4.5.3 *Types of SDOs' deliverables and their content*

Each SDO also has its own vocabulary for referring to different types of standards according to their content. For instance, ISO and CEN-CENELEC, distinguishes between different deliverables which each have a different normative power: International standards (IS) – respectively European Standards (EN), Technical Specifications (TS), and Technical Reports (TR), among others. TSs concern work that is still under development, but for which it is believed that there will be no agreement on an IS – respectively an EN – in the near future. They are published to obtain feedback with a view to the final objective of publishing an IS. On the contrary, the content of TRs is very different from that of ISs and TSs. They are more informative and generally describe the current state of the art.

In addition, IS or EN standards, unlike TRs which are merely informative, include requirements – expressed as "shall" – and guidelines – expressed as "should". Compliance with each requirement is mandatory to comply with the standard, while compliance with the guidelines is recommended, but may depend on the individual case. In this thesis, we refer to all these documents as "standards", even though they may fall into different categories of standards[43].

4.5.4 *Link with certification and accreditation*

Standardisation cannot be understood in isolation. It is part of a wider framework, including certification and accreditation, which is sometimes referred to as the *"tripartite standards regime"* (Busch, 2010). Within this framework, standardisation bodies develop technical standards, certification bodies provide assurance that a product or a set of processes comply with these standards, and accreditation bodies attest to the competence and independence of certification bodies in carrying out these activities.

While certification is primarily used to demonstrate conformity to standards,

[37] This is the case for instance at IEC (IEC, n.d.). But ISO makes a further distinction between: "full members" who can participate in standard setting and adopt and sell standards; "correspondent members" who only participate as observers; and "subscriber members" who can simply be kept up to date with standards work (ISO, n.d.g).

[38] There are some exceptions to this. For instance, ETSI and ITU membership systems work differently than the other SDOs: stakeholders can participate directly in standardisation work, without registering first in a NSB (ETSI, n.d.; ITU, n.d.).

[39] The seven other members are the United Kingdom, the Republic of North Macedonia, Serbia and Türkyie, plus three countries of the European Free Trade Association (Iceland, Norway and Switzerland). See members' list at (CEN, n.d.a; CENELEC, n.d.).

[40] In Germany, the DIN participates in CEN and the DKE participates in CENELEC. They are both present in JTC 21 on AI.

[41] Such as the European Trade Union Confederation (ETUC).

[42] Such as the European Association for the Co-ordination of Consumer Representation in Standardisation (ANEC).

[43] We present our categorisation of standards in Chapter 5, Section 5.5.1. There is great overlap between our categorisation and the types of SDOs' deliverables. For example, TRs, which only describe the state of the art, will necessarily be "information standards".

its is ultimately a way for an organisation to show consumers, regulators or business partners, that it has implemented the necessary safeguards when developing a product (van Leeuwen, 2017). Once the product or process has been successfully assessed, the certification body generally issues a certificate, seal or mark to demonstrate compliance. Most of the time, certification is carried out by third party and is voluntary. However this is not always the case: CE marking in Europe, for example, can be seen as a mandatory form of certification, often carried out by the manufacturers themselves[44]. Although standards from "recognised" bodies, such as ISO, may be used for certification, other means of assessment can also be used. For instance, labels are also considered to be a form of private certification, where the certification scheme is developed by a company which also conducts the audit and issues the certificate. In these cases, the certification audit is based on their own criteria, which are often kept secret.

The third layer of this framework is accreditation, which Kamara (2021) describes as *"certifying the certifiers"*[45]. Accreditation aims at ensuring that certification bodies, and specifically conformity assessment bodies in Europe who deliver certificate for CE marking, are suitable to perform the relative tests and audits leading to the deliverance of a certification. In the EU, rules for accreditation are given in Regulation 765/2008 (European Parliament and Council, 2008b). The regulation establishes that each Member state must have a National Accreditation Body (NAB). The criteria for accreditation are generally based on an entity's internal competencies and skills in the field in which it carries out certification activities, and on its independence from the organisations it aims to certify. In addition, accreditation bodies such as the NABs are also generally subject to an assessment, which may be carried out by other accreditation bodies[46].

But the three levels of the tripartite regime are always interacting and are not as separate as one might think. Firstly, although certification bodies are generally separate from standardisation bodies, many standardisation bodies, especially at national level, have a certification branch dedicated to selling an auditing service and issuing a certificate for the standards they also sell[47]. Secondly, as accreditation is itself a form of certification, accreditation assessments are often based on standards. For instance, ISO/IEC 17021-1 (2015) can be used for the accreditation of bodies providing certification of management systems, and ISO/IEC 17011 (2017) can be used to control accreditation bodies themselves. In the case of AI, ISO/IEC 42006 (n.d.d) could be used for the accreditation of bodies providing certification for ISO/IEC 42001 (2023b), i.e. the certification of management systems applied to AI. Finally, since an accredited certification body is supposed to be less prone to fraudulent activities, according to Kamara (2021), there is an interconnection between accreditation and the reliability of compliance with a standard.

4.6 *Product safety in the EU*

4.6.1 *A short history of product safety in the EU*

The mid-1980s is often seen as a turning point in European legislation, as the stagnation of the internal market and criticism from European bodies intensified (Senden, 2005). The European Commission's aim was to establish a *"new legislative culture"*, with the ambition of achieving greater flexibility and greater diversification of regulatory instruments. This phenomenon only intensified in the late 1990s

[44] See Section 4.6.4.

[45] A list of accredited certification bodies can be found at (IAF, n.d.).

[46] Regulation 765/2008 calls this process *"peer evaluation"*.

[47] In France for instance, the national body AFNOR provides this types of services thought its "AFNOR Certification" branch: `https://certification.afnor.org/en`. Note however, that ISO does not directly provide certification services.

and early 2000s, leading to the introduction of self-regulation and co-regulation mechanisms (Senden, 2005). Co-regulation mechanisms are notably present in the European approach to product safety. Indeed, as of today, manufactured products sold in the European Union must comply with various health and safety regulations before they can enter the EU market. This product safety framework ensures that all products of the same type sold in the EU meet the same safety levels. Harmonising safety requirements facilitates the exchange of goods and improves the internal market.

This way of working dates back to the 1980s, when the European Union established a new framework to assess conformity of products, known as the "New approach" (European Union, 2011). The New Approach consisted of a series of Council Resolutions and Decisions on technical harmonisation and standards (European Council, 1985b), conformity assessment (European Council, 1989), and CE marking (European Council, 1993). Under the New Approach, a large number of Directives was adopted: they were applicable to specific product categories for which self-certification was needed. The certification consisted notably of the application of CE-marking on compliant products. There were 22 product Directives in total (Paul and Loh, 2023), ranging from medical devices, to toys, electrical products and machinery. The main change brought about by the New Approach, compared to the Old Approach, was to not define technical specifications directly in the directives, but rather to set out the "essential requirements" that products should meet, leaving manufacturers[48] more flexibility as to the means of achieving compliance (CEN, 2019). Essential requirements define the results to be attained but do not specify the technical solutions for doing so (European Commission, 2022d). This approach gave an important role to technical standards and the bodies responsible for the assessment of conformity, the so-called "notified bodies".

This framework was later replaced by the New Legislative Framework (NLF) in 2008, which takes much of its core from its predecessor. It is a package of measures (European Parliament and Council, 2008b,a, 2019) that aim to improve market surveillance to better protect consumer from unsafe products, and improve conformity assessments by laying down stronger and clearer rules for conformity assessment bodies (European Commission, n.d.r). It notably clarifies the use of CE marking and creates a toolbox of measures for use in product legislation, including definitions of terms and procedures. Indeed, for a product covered by a NLF legislation to enter the European market, it must be CE marked[49]. CE marking has a dual use: it allows consumers to benefit from the same level of – presumably – high protection throughout Europe and allows the free movement of products within Europe by harmonising legislation. Products bearing the CE mark can be traded in Europe without restrictions (European Commission, n.d.c). Before development of the CE mark, trade was limited by differences in national product requirements between member states (Hanson, 2005).

In addition to the official regulations and directives which constitute the NLF, the EU has also published a Blue Guide (European Commission, 2022d), intended to be the reference document for all information on the operation of the framework, particularly with regard to standards, conformity assessment and CE marking. It can therefore be used by companies to familiarise themselves with the way the NLF works.

[48] While the European Commission (n.d.q) usually prefers the term "manufacturer" for product safety legislation, the AI Act uses the term "provider", defined in Article 3(3). We will use the former when discussing product safety legislation in general, and the latter when discussing the AI Act specifically.

[49] CE marking is applicable throughout the European Economic Area (EEA).

4.6.2 *Types of standards in the EU*

European and international standards

There are many ways of defining what a standard is and Europe in particular has its own definition. Indeed, in Regulation 1025/2012 (European Parliament and Council, 2012), the European Union makes the distinction between mere *"technical specifications"*, and *"standards"*. While a technical specification is defined as *"a document that prescribes technical requirements to be fulfilled by a product, process, service or system"* (art 2(4) Reg. 1025/2012), a "standard" is *"a technical specification, adopted by a recognised standardisation body"* (art 2(1) Reg. 1025/2012).

Why not choose the EU definition of a "standard"?
The definition of "standard" adopted in this thesis, and given in Section 4.2.1, differs from the European Union's definition. Indeed, the EU definition, like the ISO definition presented in Section 4.2.1, focuses on documents drawn up by "recognised" organisations. Although many of the standards we discuss in this thesis, such as CEN-CENELEC and ISO/IEC standards, meet this definition, we did not want to limit ourselves to these documents. Indeed, for AI, many other organisations, such as IEEE, are also proposing standards that deserve to be analysed. Additionally, not all documents developed by "recognised" SDOs are designated "standards" by these organisations. "Technical reports" or "technical specifications" that are relevant for our study can also be published by SDOs but are not considered standards by ISO/IEC or CEN-CENELEC. All of these various initiatives are included in our study, and we discuss them in Chapter 6 and Chapter 5 in particular. As a result, we have chosen a broader definition of "standard" that resembles the definition of "technical specifications" in Regulation 1025/2012, in order to encompass more types of technical documents and broaden the scope of the research.

Regulation 1025/2012 goes further by also defining *"international standards"*, adopted by international standardisation bodies, *"European standards"*, adopted by European bodies, and *"national standards"*, adopted by national bodies. The SDOs developing *"international standards"* in accordance with Regulation 1025/2012 are ISO, IEC and ITU, and the SDOs developing *"European standards"* are the ESOs, i.e. CEN, CENELEC and ETSI[50]. These various standards-related term are represented in Figure 4.2, as defined by Regulation 1025/2012.

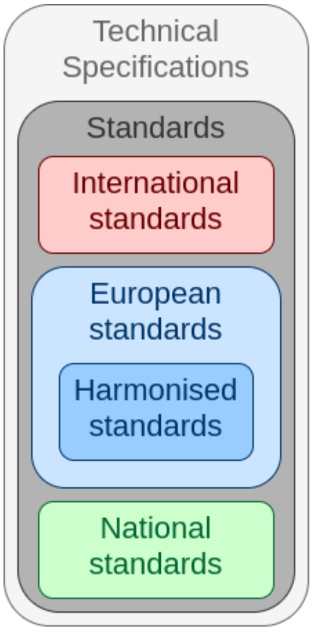


Figure 4.2: Types of standards, as defined by Regulation 1025/2012.

Harmonised standards

The ESOs may also adopt a particular type of European standard, called a *"harmonised standard"*, which play a major role under the NLF. A harmonised standard (hEN)[51] is defined in Regulation 1025/2012 as:

> ***Definition***
>
> **Harmonised standard (Reg. 1025/2012).** *European standard adopted on the basis of a request made by the Commission for the application of Union harmonisation legislation.*

[50] See Section 4.5.1 for an overview of the ESOs and international SDOs.

[51] See the box below for more information on the acronym.

Why choose the acronym "hEN"?

Many acronyms have been used by the academic literature to designate these standards: "HSs" for Harmonised Standards, "HESs" for Harmonised European Standards, or "HTSs" for Harmonised Technical Standards. In this thesis, we chose the acronym hEN which is used by ESOs, such as CEN-CENELEC. The letters EN are placed in front of the name of a standard to indicate that it has been adopted by the ESOs and is therefore considered to be a European standard. The letter h is added to indicate that it is additionally a harmonised standard. Choosing the acronym "hEN" to designate harmonised standards allows us to use a vocabulary closer to that of the practitioners and creators of these standards.

The "standardisation request" (SR) delivered by the EU Commission to the ESOs acts as a mandate, whereby the Commission officially ask the ESOs to deliver the standards. The standardisation request notably includes details on the scope, timelines, and legal requirements that the requested standards should fulfil, called "SR items". These items correspond to the essential requirements of the law which must be completed by technical standards. hENs are therefore intended to support a specific European directive or regulation by specifying how the requirements of the law can be verified by means of technical criteria. As such, they can help manufacturers or notified bodies to assess conformity.

hENs can be published in the OJEU and thus acquire special legislative powers. Indeed, products manufactured in accordance with hENs benefit from a "presumption of conformity". This means that the essential requirements covered by hENs are presumed to be automatically met if the products comply with that standard. Manufacturers may then benefit from simplified conformity assessment procedures (Hernalsteen and Kohler, 2022). According to the ESOs, *"legislation and policy supported by the use of standards have a high level of stakeholder acceptability, can minimise burdens on industry and can reduce the cost of compliance"* (CEN-CENELEC, 2015a). Additionally, the presumption of conformity afforded by hENs encourages their adoption and avoids legal claims concerning hENs when a manufacturer's position on the market is affected by these standards (Schapel, 2013). But this quasi-legislative role also brings with it its share problems, notably regarding access and copyrights[52]. However, not all hENs are published in the OJEU and benefit from this superior legal status. Some standards might be requested by the European Commission to address standardisation gaps, without supporting a specific legislation (Hernalsteen and Kohler, 2022).

The standardisation system put in place by the EU therefore relies on the strong collaboration between the ESOs and the European Commission, the Commission being the legislative body and the ESOs being an executive body.

4.6.3 Process for the harmonisation of technical standards

As harmonised standards are intended to support European law, contrary to "traditional" European standards, their adoption process does not solely depend on the ESOs. The European Commission intervenes at the beginning of the process to formalise the request, and at the end to validate the final version of the standard, through a harmonised standards consultant (HAS consultant). The steps of the harmonisation process are represented in Figure 4.3.

[52] For a discussion on the place of hENs in EU law, see Chapter 7, Section 7.3.

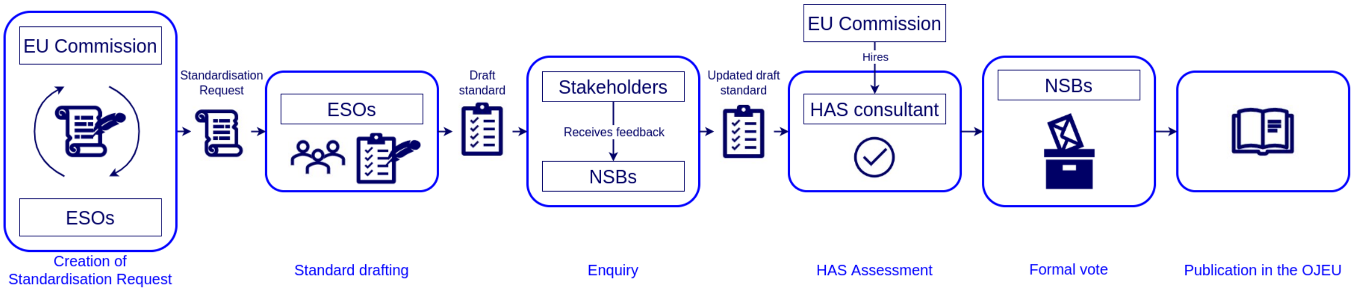


Figure 4.3: Process for the harmonisation of technical standards, adapted from (Pouget, 2024).

First, the Commission has to propose a standardisation request to the ESOs. The standardisation request can be addressed to one or several ESOs specifically. Once the request is published, the ESOs have the choice to either accept or reject it. However, the ESOs do not find out about the request when it is published: they contribute directly to its development. In fact, since 2016, a mechanism has enabled the European Commission and CEN-CENELEC to discuss standardisation requests before official approval, in order to avoid too many requests being rejected (CEN et al., 2021). These negotiations take place within Standardisation Requests Ad-hoc Groups (SRAHG) (CEN-CENELEC, 2015b).

Once the request has been accepted, the standard is developed as usual by the ESOs[53]. After the first draft is finished, the process enters the phase of "enquiry" in which the NSBs are responsible for gathering feedback from any relevant stakeholders and updating the draft accordingly.

HAS consultants are selected and hired by a HAS contractor who acts as the formal point of contact between the European Commission and the ESOs. Currently, the Commission has a contract with Ernst & Young (Pouget, 2024) to provide these consultants. The HAS consultant is tasked with carrying an assessment of the standard in question, using a checklist, and compiling the results in a "HAS Assessment report" (CEN, n.d.c). The checklist is used in particular to verify whether the scope and provisions of the standard cover the request, and whether the terminology remains consistent with the legislation[54]. The HAS consultants can also be included throughout the process to prevent the case of non-compliance. Indeed, if the standard is not deemed compliant, its reference cannot be published in the OJEU unless the standard is adapted[55]. If the standard is compliant, then the process moves to the next phase: the formal vote by the NSBs. Once the standard was found compliant by the HAS consultant and adopted by the vote of the NSBs, its reference can be published in the OJEU.

4.6.4 *Conformity assessment and CE marking*

hENs are used in particular during the conformity assessment procedure, a necessary step to place a product on the EU market, in which the product is tested to check that it complies with the applicable requirements. For some applications, the conformity assessment is carried out by the manufacturer themselves or a subsidiary, while for others it is necessary to call on the services of a notified body. For some products, the procedure includes a control of the quality management system, in addition to the technical checks.

Manufacturers are free to choose any standard, whether hENs or other technical

[53] In this thesis, we do not describe in detail the different steps of the development and adoption of a standard within the ESOs. However, we provide some information on the functioning of these processes, as explained by standardisation experts themselves, in Chapter 8, Section 8.4.3.

[54] The checklist can be downloaded from this page: (CEN, n.d.c).

[55] In some cases, the reference can be published in the OJEU, but with a notice.

specifications, including some they have defined themselves. But in the end, they must be able to prove that these technical specifications are sufficient to ensure compliance with the legislation[56].

Alongside the technical tests, the manufacturer must draw up technical documentation and make it available to the national authorities. Once the conformity assessment is completed, manufacturers must then affix the CE mark to the product and draw up the declaration of conformity, which should be presented to national authorities upon request[57]. Manufacturers are responsible for CE marking (European Commission, n.d.q). By affixing the CE mark to their product, they declare that they have ensured – through a conformity assessment – that the product meets the essential requirements of the applicable European legislation. An example of a conformity assessment procedure is shown in Figure 4.4.

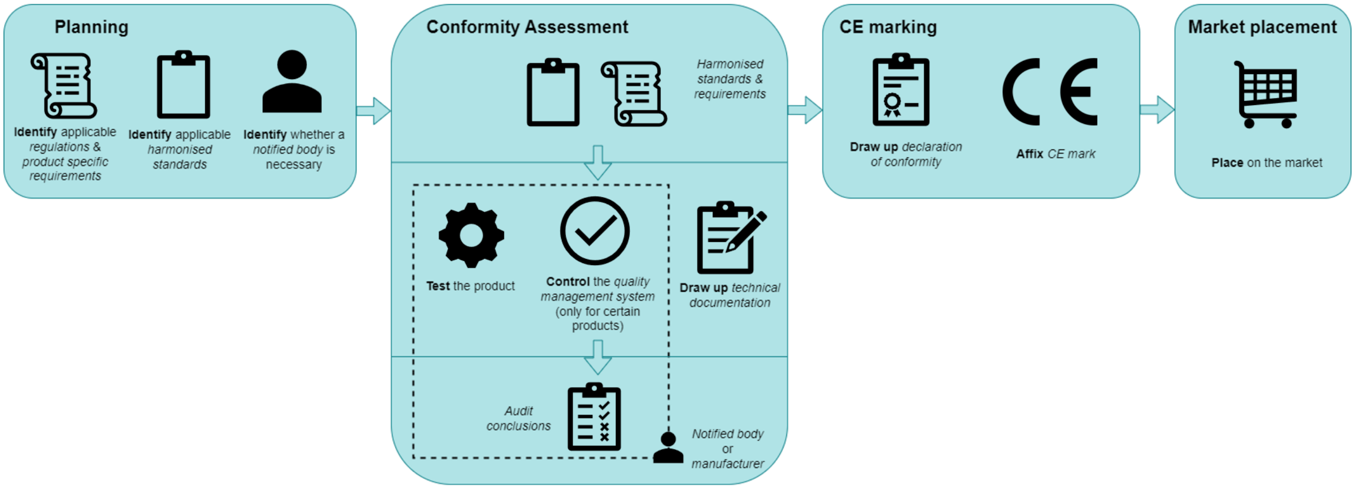

Planning
Identify applicable regulations & product specific requirements
Identify applicable harmonised standards
Identify whether a notified body is necessary
Conformity Assessment
Harmonised standards & requirements
Test the product
Control the quality management system (only for certain products)
Draw up technical documentation
Audit conclusions
Notified body or manufacturer
CE marking
Draw up declaration of conformity
CE
Affix CE mark
Market placement
Place on the market

Figure 4.4: Conformity assessment of a product under the NLF. All the steps are carried out by the manufacturer, except those between the dotted lines, which may also be carried out by a notified body depending on the type of product. Representation from the author, using information from (European Commission, n.d.q).

Today, the product safety framework defined by the NLF is used for a wide range of products in the EU, including AI since the adoption of the AI Act.

4.6.5 *European governance for standardisation*

To ensure that the European standardisation strategy succeeds, Europe has implemented a full ecosystem to supervise activities and provide funding. In particular, a specific unit at the Joint Research Centre (JRC) of the European Commission supervises standardisation efforts and works closely with standardisation experts (European Commission, n.d.v). In the case of AI, the AI Act will rely on a whole European ecosystem to implement the text, and has notably created the AI Office[58]. A team within the AI Office's Unit CONNECT.A.2 on "AI Regulation and Compliance"[59] will therefore be responsible for overseeing specific AI standardisation efforts, in conjunction with the JRC. Additionally, the 2022 European Commission standardisation strategy has created a new position, the Chief Standardisation Officer, responsible for steering and supervising the various standardisation activities across the Commission (European Commission, 2022e).

But the Commission can also rely on advisory groups, including the European Multi-Stakeholder Platform on ICT Standardisation (MSP), set up in 2011 (European Commission, 2011a), and the High-Level Forum on Standardisation (the Forum) set up in January 2023 (European Commission, 2022a). While the MSP advises on *"all*

[56] For more information on the difference between using harmonised standards and other standards during conformity assessment, see the discussion in Chapter 7, Section 7.3.

[57] The 6 steps to affix CE marking are presented in (European Commission, n.d.q).

[58] For more information on this ecosystem and the AI Office, see Chapter 3, Section 3.5.4.

[59] See the structure of the AI Office in Chapter 3, Figure 3.7.

matters related to the implementation of ICT standardisation policies" (European Commission, n.d.g), the Forum specifically aims at *"identifying standardisation priorities"* and *"bring more alignment between European policy priorities [...] and standardisation actions"* (European Commission, n.d.m).

In addition, given that standardisation work is voluntary and that many experts are not funded, programmes have been launched by Europe to support them. The most prominent of these programmes, the ICT Standardisation Observatory and Support Facility in Europe (StandICT.eu), provides funding to support the participation of European experts in standardisation initiatives in the Information and Communications Technology (ICT) sector[60]. Across a series of nine calls from May 2023 to April 2025, StandICT will provide nearly €3 million to standardisation experts (StandICT.eu, n.d.). An individual may apply and receive funding more than once, but up to a maximum of €60,000 in total. Experts can apply regardless of their affiliation but between 85% and 90% of the funds go either to SMEs, academia, or IT consultants [61]. The SDOs targeted are diverse, but ISO/IEC experts receive the most funding, followed by CEN-CENELEC experts. All the funds are shared among all ICT fields, but AI leads the way in terms of the total number of applications funded, with around 10 experts receiving funds at each new call.

4.7 The AI uses product safety tools

4.7.1 The AI Act has a risk-based approach

The AI Act (European Parliament and Council, 2024b) is a risk-based regulation where AI systems are classified based on their level of risk[62]. "Risk" is understood in the text as the *"combination of the probability of an occurrence of harm and the severity of that harm"*[63], as stated in the Article 3(2) of the AI Act. For certain AI systems, including chatbots and deepfakes[64], only transparency requirements apply (Art. 50) and for other systems that are found to be very low risk, no regulatory burden applies. On the contrary, systems presenting an unacceptable risk are prohibited entirely (Art. 5). This is the case of social scoring and certain facial recognition applications. In addition, successive amendments to the AI Act have introduced the notion of General Purpose AI (GPAI) models which are trained using large amounts of data and can be used in a variety of downstream tasks. These systems are subject to additional documentation requirements (Art. 53), and if they are considered to present a "systemic risk", i.e. if they have significant capacities generally exceeding a certain complexity threshold, they are subject to additional assessment requirements (Art. 56)[65].

But the core focus of the AI Act is on "high-risk AI systems". High-risk systems can be products covered by harmonised legislation presented in Annex I, including medical devices, machinery or toys. They may also be systems operating in the fields of application listed in Annex III (Art. 6(2)). They are then called *"standalone AI systems"* (rec. 52) and include for instance biometric systems, AI for essential services, education, workplace, law enforcement, border control, justice and so on. The list of high-risk systems can be amended by the Commission, if a new use case is found to create high risks (Art. 7(1)). In addition to these systems, it should be noted that GPAI models can be implemented in high-risk applications and therefore also fall under this category. Systems that are considered high-risk must comply with the requirements set forth in Chapter III, Section 2, in relation to

[60] StandICT.eu itself has received funding from Horizon Europe (European Commission, 2020b), a *"research and innovation funding programme"* which initially ran from 2014 to 2020 – called Horizon 2020 (European Commission, n.d.n), and has been renewed for the period 2021 to 2027, with a total budget of €93.5 billion (European Commission, n.d.o).

[61] Based on the statistics provided for the last 4 calls (StandICT.eu, 2023a,b, 2024a,b).

[62] See Chapter 3, Section 3.5 for more information on the AI Act structure and requirements.

[63] Note that a similar definition is given in the General Product Safety Regulation (European Parliament and Council, 2023c).

[64] Chatbots are AI systems that generate textual content for interaction with a natural person. Deepfakes are images, videos or audio files generated by an AI that generally reproduce existing individuals in imaginary scenarios.

[65] For more information on the different categories of AI systems, see Chapter 3, Section 3.5.2.

risk management, data and data governance, technical documentation, record keeping, transparency and provision of information to users, human oversight, accuracy, robustness and cybersecurity. Within those requirements, risk management is a key element, particularly when AI is used in high-stakes situations (Schuett, 2023b). Providers of high-risk AI systems must establish, implement, document and maintain a risk management system, consisting notably of the identification of known and foreseeable risks, as well as adoption of appropriate measures to eliminate or mitigate those risks (Art. 9). Residual risks must be reduced to a "reasonable" level, dictated by the state-of-the-art (Fraser and Bello y Villarino, 2023).

4.7.2 CE marking will show that AI systems comply with the regulation

The AI Act is a product safety regulation under the NLF. As such, it establishes an ex ante accountability framework for AI (Castets-Renard and Besse, 2022) in which proof of compliance with general requirements is a prerequisite for the *"placing on the market or putting into service"* of AI systems (Art. 2). In particular, the AI Act stipulates that high-risk AI systems must undergo a conformity assessment procedure and, when they are found to be compliant, providers must draw up an EU declaration of conformity and affix the CE mark on the product (Art. 16). The conformity assessment procedure is carried out either by a third party or by the provider of the AI system, depending on: (i) if the system falls under an application use case listed in Annex III; and (ii) if the provider has applied hENs (Art. 43).

In most cases, the conformity assessment will be carried out by the providers themselves and the result of the conformity assessment will not be made public. This procedure, based on internal control, is described in Annex VI and simply consists of providers assessing compliance with the requirements[66] using the information contained in the technical documentation[67]. This procedure will be applied to "stand-alone" high-risk AI systems covered by Annex III, with the exception of biometric systems (Art. 43(2)). On the other hand, a third-party audit will be required for high-risk AI systems covered by the NLF regulations listed in Annex I (Art. 43(3)) and for biometric systems if hENs do not exist or if the provider has not applied them (Art. 43(1)).

The third party audit will be conducted by a notified body, i.e. a conformity assessment body "notified" by a national authority (Art. 3(22))[68]. These notified bodies are mainly private entities, designated by a EU country to conduct conformity assessments on a certain range of products (European Commission, n.d.s)[69]. Even before the AI Act, notified bodies were used to assess the compliance of products with NLF regulations. It should be noted however that for AI systems, the procedure for the conformity assessment carried out by a notified body, described in Annex VII, mostly require a verification of the technical documentation drawn up by the provider. The notified body may require additional tests to be carried out by the provider to complete the documentation (Annex VII.4.4). Access to training and testing data will also be possible but only when necessary (Annex VII.4.3), and access to the model itself will only be possible when *"all other reasonable means to verify conformity have been exhausted"* (Annex VII.4.5). Following the conformity assessment, the notified body issue a certificate stating that the high-risk AI system meets the requirements of the AI Act (Annex VII.4.6).

Recital 125 recognises that, ideally, conformity assessment should be carried out

[66] Providers are also required to verify the quality management system (Art. 17) and the post-market monitoring systems (Art. 72).

[67] The technical documentation will itself be drawn up by the provider.

[68] Called a "notifying authority" (Art. 3(20)).

[69] For a complete list of all notified bodies, see (European Commission, n.d.t).

by a third party. However, in the same recital, the Commission justifies the choice of restricting the use of notified bodies to products already covered by NLF regulations by explaining that this ensures that the notified bodies are already experienced certifiers. The Commission reserves the right to amend this provision in the future, and to subject the high-risk systems in Annex III to a third-party audit, in the light of the effectiveness of internal control and the resources needed to cover all such systems (Art. 43(6)). However, this scenario is unlikely to occur in the near future (Wachter, 2024).

4.7.3 Harmonised standards will provide a technical means of assessing compliance

hENs are, in the AI Act, seen as a way to operationalise mandatory requirements (European Commission, 2021e, Explanatory Memorandum, section 2) while reducing costs (Explanatory Memorandum, section 2.3). Recital 121 of the AI Act further states that *"standardisation should play a key role to provide technical solutions to providers to ensure compliance"*. hENs will notably simplify conformity assessments, particularly for systems which do not require third-party audit. Article 40 of the AI Act reminds us of the power of hENs: high-risk AI systems or general-purpose AI models which are found to be in conformity with hENs published in the OJEU will be presumed to be in conformity with the corresponding requirements set out in the AI Act. This means, for example, that providers of certain high-risk AI systems can opt out of a third-party conformity assessment and fully rely on internal control, if they choose to apply hENs (Art. 43(3)). If they choose not to apply hENs, they must demonstrate by other means how the specifications they use permit products to comply with the essential requirements (European Commission, 2022d, p.55), a more challenging task than if they simply applied a hEN. Some experts therefore believe that it is in standardisation that the real rule-making will occur (Veale and Borgesius, 2021).

The European Commission issued a standardisation request to the ESOs regarding standards for the AI Act which was formally published as a Commission Implementing Decision in March 2023[70] (European Commission, 2023a). In the request, the Commission asks specifically CEN and CENELEC to cover ten subjects related to the requirements for high-risk systems[71]. These topics correspond to the requirements for high-risk AI systems set out in Chapter III, Section 2 of the AI Act. A second version of the standardisation request should be published by the Commission before the end of 2024. In particular, it should be aligned with the latest version of the AI Act, which has now been officially published[72]. The final version of the AI Act provides for the use of hENs to cover requirements other than those for high-risk AI systems, in particular for the requirements of the GPAI models listed in Article 50. It is not yet known whether the new standardisation request will require ESOs to cover these aspects.

CEN and CENELEC are now working on hENs for the ten topics listed by the current request, as well as other topics, at their own discretion. In particular, the first standard that is being developed specifically for harmonisation purposes is the "AI Trustworthiness framework", a single wide-ranging standard which aims to cover various items of the standardisation request (ETUC, 2024)[73].

[70] A draft version was released publicly in 2022 (European Commission, 2022f).

[71] Risk management system for AI systems, governance and quality of datasets used to build AI systems, record keeping through logging capacities by AI systems, transparency and information provisions for users of AI systems, human oversight of AI systems, accuracy specifications for AI systems, robustness specifications for AI systems, cybersecurity specifications for AI systems, quality management system for providers of AI systems, including post-market monitoring process, and conformity assessment for AI systems.

[72] The first standardisation request dates back to an earlier version of the AI Act.

[73] See the discussion on the ongoing work on the Trustworthiness framework and other standards in Chapter 8, Section 8.5.3.

4.8 Conclusion

In this chapter, we have provided background on technical standards and the new legislative framework for product safety in Europe. We have shown that the role of standards has evolved over recent decades, becoming more horizontal, focusing more on quality management than on the actual technical properties of systems, and playing a more legal role than before, particularly in the EU where harmonised standards are used to support European legislation. We presented the main standardisation organisations and their functioning, where experts in standardisation must first register in a national body to participate in standardisation efforts. The six main SDOs – at international level, ISO, IEC and ITU, and at European level, CEN, CENELEC and ETSI – are often collaborating and adopting each others standards. They also have their own vocabulary for talking about standards, as standards with different content go by different names. However, in the context of this thesis, we have chosen to retain a broader definition of the term standard, which is not influenced by who writes it or what it contains, but rather by its purpose: to capture industry expertise.

In addition, we have examined the place of harmonised standards and CE marking in the European AI Act. We have shown that the AI Act strongly relies on harmonised standards to assess compliance with the requirements for high-risk AI systems. Moreover, the conformity assessment is usually conducted by AI systems providers themselves, the involvement of a notified bodies being reserved for systems that are already covered by harmonised legislation under the NLF, and not new applications of AI systems. The development of standards and compliance check procedures could then be left mostly to private actors.

PART II

A FOCUS ON STANDARDS TO REVEAL TENSIONS BETWEEN REGULATORY MECHANISMS

Chapter 5
Standardising ethics and fundamental rights: the difficult task of AI standards

Where does this chapter come from?

This chapter extends on the following articles:

> Gornet, M. and Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, 13(3). `https://doi.org/10.14763/2024.3.1784`

> Gornet, M. and Maxwell, W. (2023b). Normes techniques et éthique de l'IA. In *Conférence Nationale en Intelligence Artificielle (CNIA @ PfIA 2023)*, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). `https://pfia23.icube.unistra.fr/conferences/cnia/publications/Gornet.pdf`

It notably reuses one sub-section and one main section of the first article[1]. It also reuses sections of the second article[2]. These sections were translated from French to English, and updated to reflect the current state of the art of standardisation initiatives.

The mixing of these two articles has one major consequence: it brings together our study of "ethical standards" and "fundamental rights standards" into a single study. We believe merging the two is possible because the two types of standards raise the same questions and problems. Subsequently, in this chapter and in the rest of the thesis, we will discuss standards that encompass broad ethical, social and fundamental rights issues. Of course, in certain sections, we will address ethical standards and standards on fundamental rights separately, where necessary.

Chapter's abstract

With the AI Act, Europe is the first region in the world to adopt a mandatory framework to regulate AI. Inspired by product safety regulation, high-risks AI systems will have to go through a conformity assessment procedure before being placed on the market. Harmonised technical standards (hENs) will notably be used to facilitate the procedure, while a CE mark will be affixed to the product to show compliance with the law. The AI Act is not the first law to rely on hENs and CE marking, European regulations have relied on them for many years to ensure product safety across a wide array of items. To date, however, they have never been

[1] The sub-section initially named "Harmonised standards were not initially designed to cover fundamental rights", and the section "Fundamental rights and technical standards".

[2] The section initially named "Une multitude d'acteurs et d'initiatives en matière de normes éthiques pour l'IA".

used to attest to compliance with fundamental rights, something the AI Act aims to achieve. In this chapter, we show how these product safety regulatory techniques have been expanded to cover protection of fundamental rights and we look at standardisation initiatives which are currently trying to address ethical, societal and fundamental rights issues. We show that unlike compliance with product safety standards, compliance with fundamental rights cannot be certified through use of technical standards because violations of rights are too context-specific and require a judicial determination. However, technical standards have an important role to play in encouraging best practices in AI governance.

5.1 Introduction

A new European regulation on AI was adopted by European institutions this year: the AI Act (European Parliament and Council, 2024b). The AI Act takes a risk-based approach to AI regulation, with AI systems falling into different categories according to the risks they pose to the *"heath, safety and fundamental rights"*. The AI Act places particular emphasis on the high-risk category, for which systems will have to undergo a conformity assessment. During this assessment, providers of high-risk AI systems themselves, or external auditors, will test the system's compliance with various requirements and obligations which are set out in the AI Act. To this end, they will refer to technical standards and, where they exist, to harmonised standards (hENs), a particular type of standard developed in Europe and which can have certain legal effects. Successful conformity assessment enables AI systems to bear the CE mark, which means they can then be distributed on the European market. The CE mark and hENs are long-established European regulatory tools to deal with product safety and already apply to a wide range of products.

However, unlike other product safety regulations, the AI Act is not only intended to protect against risks to safety, but also against adverse effects on fundamental rights. Consequently, hENs and CE marking could also apply to the protection of fundamental rights. In addition, the strong influence of ethics charters on the AI Act could also push standards to focus on integrating these aspects. Standardising ethics or fundamental rights is not entirely new. Standards on fundamental rights were first introduced with the General Data Protection Regulation (GDPR) (European Parliament and Council, 2016b) and a few initiatives on standardising ethics have emerged in recent years. However, with the AI Act, the extension of the product safety approach to fundamental rights takes the stakes of technical standards to a new level. In particular, this raises the question of how – and if – technical instruments such as standards can address these normative issues.

In this chapter, we begin in Section 5.2 by recalling the dual discourse of the AI Act, which aims both to protect the fundamental rights of EU citizens and to improve the internal market through a product safety approach based on compliance with technical standards. We show that technical standards are not intended to cover these issues, but will be forced to do so by the AI Act. In Section 5.3, we provide an overview of all the current initiatives to standardise AI ethics and fundamental rights, in Europe and beyond. Indeed, it is necessary to have a global view of standards, because work carried out outside Europe can be used to assess compliance with European law[3], but will also influence market dynamics. In Section 5.4, we then dive into the problems with addressing normative questions

[3] We explain the collaboration of European and international organisations in Chapter 4, Section 4.5.1 and discuss the sovereignty issues this can raise in Chapter 7, Section 7.5.

through technical standards. We show that technical and normative issues are always intertwined. But when technical choices are seen as neutral or taken out of context, there is a good chance that compliance to these standards will lead to ethics washing practices. Finally, in Section 5.5, we present our own classification of the types of standards being developed today and give our views on the prospects for future standardisation initiatives. We propose that standards should at least avoid the strong normative choice of setting thresholds for technical criteria which directly affect fundamental rights.

5.2 *The AI Act takes standards into the realm of fundamental rights*

5.2.1 *Product safety in the AI Act*

The AI Act (European Parliament and Council, 2024b) is part of the New Legislative Framework (NLF) for product safety in Europe[4]. Products covered by the NLF must undergo a conformity assessment before being placed on the European market. Once the conformity assessment has been completed, the manufacturer affixes a CE mark (European Commission, n.d.c) to the product, thereby certifying that it complies with the regulation and assuming responsibility in case of non-compliance. CE marking allows products to be distributed throughout Europe, without having to go through different procedures each time the product is placed on the market in a new country. This way, consumers can be sure that every product in the same category distributed on the European market has been checked for conformity by the manufacturer. All products in the same category are therefore expected to meet the same safety standards throughout Europe. This system harmonises rules and facilitates the distribution of goods, as one of the objectives of the European Union is to have a common market where products can be traded across countries. The NLF makes this harmonisation of product safety possible and improves the internal market overall.

Under the AI Act, high-risk AI systems fall under the NLF and will have to undergo a conformity assessment and CE marking before being distributed on the market. The high-risk category is the second highest risk category in the AI Act, just after “unacceptable risk” systems, which are totally banned in Europe. The high-risk category includes products already covered by the NLF that are subject to conformity assessment under their sector-specific legislation, such as toys, machinery or medical devices. But the AI Act also lists new sectors where the use of an AI system will be considered high-risk, such as education, employment, justice, etc. For systems subject to existing sector-specific regulations, the conformity assessment will be performed by the existing authorities, known as notified bodies. But for systems used in new applications, conformity assessment will be based mainly on internal control[5].

To pass the compliance assessment, high-risk AI systems will have to comply with a number of obligations, such as transparency, data quality, human oversight, and so on, which are enumerated in Chapter III, Section 2 of the AI Act. To ensure compliance, providers of AI systems will be able to rely on harmonised standards (hENs), a specific type of technical specifications, which can have legal power in the EU if they are published in the Official Journal of the European Union (OJEU). In particular, compliance with hENs automatically implies compliance with the corresponding legal text: this is known as the presumption of conformity[6]. hENs are first

[4] For more information on the NLF, see Chapter 4, Section 4.6.

[5] See Chapter 3, Section 3.5.2 and Chapter 4, Section 4.7 for a more detailed explanation of, respectively, the different risk categories, and the different required compliance assessments under the AI Act.

[6] For more information about harmonised standards and their legal role, see Chapter 4, Section 4.6.2 and Chapter 7, Section 7.3.1.

mandated by the European Commission in a standardisation request, then drawn up by one or several of the European Standardisation Organisations (ESOs)[7]: the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardisation (CENELEC) and the European Telecommunications Standards Institute (ETSI). But beyond harmonised standards, any type of technical instrument can be used to demonstrate compliance with the law[8].

5.2.2 *Trustworthiness, ethics and fundamental rights in the Act*

As regard to fundamental rights, the Commission's explanatory memorandum presented just before the text of the first proposal of the AI Act and constituting an important aid to the interpretation of the legislation, contains a list of rights whose protection should be enhanced by the AI Act (European Commission, 2021e, Explanatory Memorandum, section 3.5). It includes, for example, the right to human dignity, respect for private life and protection of personal data, non-discrimination, equality between women and men, freedom of expression, freedom of assembly, right to an effective remedy and to a fair trial, rights of defence and the presumption of innocence, the general principle of good administration, etc. These rights must be protected in particular because of the *"specific characteristics (e.g., opacity, complexity, dependency on data, autonomous behaviour) [of AI systems that] can adversely affect a number of fundamental rights"*. Some of the requirements for high-risks AI systems can also be considered to be revisited versions of fundamental rights, such as human oversight or risk mitigation mechanisms (Almada and Petit, 2023). For some authors, fundamental rights are the "backbone" of the AI Act, placed for the first time in product safety regulation on an equal footing with safety and health protection (Almada and Petit, 2023). On the contrary, for others, they are just an "afterthough" in a law that is primarily focused on market incentives (Castets-Renard and Besse, 2022). Some organisations therefore regret that fundamental rights are treated as a "buzzword" in the AI Act (EDRi, 2024).

Beyond fundamental rights, the Commission insists on its desire to integrate ethical considerations into the supervision of AI systems. In the explanatory memorandum, the European Commission states that the proposed essential requirements are inspired by the "Ethics Guidelines" of the High-Level Expert Group on AI (HLEG, 2019a). In particular, some of the language elements of the HLEG are reflected in the AI Act (European Parliament and Council, 2024b), such as the notion of "trustworthiness"[9], a broad concept that incorporates ethical values and legal norms (Laux et al., 2024). Indeed, Recital 3 which states that the AI Act should ensure a high level of protection *"in order to achieve trustworthy AI"*. The seven key requirements – also called principles – set out by the HLEG are even recalled in Recital 27 of the AI Act[10]. In a previous version of the text, the Parliament even listed some *"general principles applicable to all AI systems"* (European Parliament, 2023a, amendment 213), directly taken from these seven key requirements[11].

There is a strong link between these ethical principles, which guide the AI Act, and a number of fundamental rights. For instance, *"transparency"* or *"diversity, non-discrimination and fairness"* relate to the fundamental rights to information and non-discrimination. The explanatory memorandum also states that it is in the Union's interest to *"ensure that Europeans can benefit from new technologies developed and functioning according to Union values, fundamental rights and principles"* (Explanatory Memorandum, section 1.1). Another example of how fundamental rights are

[7] The standardisation request can be addressed directly to one or several ESOs.

[8] Although this is more difficult, see in particular Chapter 7, Section 7.3.1.

[9] For more information on "trustworthiness" in the HLEG guidelines, see Chapter 3.

[10] These principles are: *"human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability"*.

[11] Except for *"accountability"*, as it was assumed that the regulation would enable this key requirement to be enforced.

taken into account can be found directly in the text of the Act: a system shall be considered high-risk if it *"pose[s] a risk of harm to health and safety, or an adverse impact on fundamental rights"* (art. 7.1(b) AI Act). This list was initially extended by the Parliament in its proposed amendments to the AI Act, which also considered harms to *"the environment, democracy and the rule of law"* (European Parliament, 2023a, amendment 246). However, these additions were mainly perceived as "rhetorical flourishes" (Smuha and Yeung, 2024), given that the rest of the text and its market-based approach did not evolve accordingly. In the end, this extension was not retained in the latest version of the text.

The AI Act also introduces in Article 27 a new mechanism to assess trustworthiness: the fundamental rights impact assessment (FRIA), inspired by the data protection and privacy impact assessments of the GDPR. FRIAs were initially introduced by the Parliament in a previous version of the text (European Parliament, 2023a, amendment 413) as their absence in the first proposition by the Commission was criticised (Edwards, 2022). A FRIA will be mandatory for high-risk systems, where the deployer is a public law body or a private entity providing public services. The FRIA will contain a list of natural persons and groups likely to be affected by the system, together with these specific risks, as well as the measures to be taken to mitigate these risks, including a description of human oversight implementation.

5.2.3 Towards standards for ethics and fundamental rights?

In its approach to regulating AI, the AI Act pushes two different dimensions. First, it is a product safety regulation based on *ex ante* compliance and which will require a conformity assessment based on harmonised technical standards[12]. Secondly, it is strongly influenced by ethical texts and has a strong fundamental rights component. As such, the AI Act pursues a dual objective of protecting individuals' fundamental rights, while enabling the free movement of data and AI systems within the Union. This is best shown by the legal basis on which the European Commission has proposed the text. Indeed, both the data protection competence enshrined in Article 16 of the Treaty on the Functioning of the European Union (TFEU) (European Commission, 2012c), and the internal market policy competence enshrined in Article 114 TFEU are cited by the Commission (Explanatory Memorandum, section 2.1). For comparison, the GDPR was only based on Article 16, while most of product safety regulation are based on Article 114. Although fundamental rights have already been addressed and protected by European law, such as in the GDPR for example, the AI Act is the first attempt to integrate fundamental rights into a product safety approach, using hENs and CE marking.

But for Almada and Petit (2023), this dual discourse means that there is a mismatch between means and ends in the AI Act: while the main objective of the AI Act is to protect fundamental rights, the product safety approach used by the legal text is not well suited to achieving this objective. For instance, the fusion of the compliance approach and the fundamental rights approach raises the question of whether standards for the AI Act will have to address fundamental rights. Similarly, as the text is heavily inspired by ethics charters and relies on several "trustworthiness" principles, what role will ethics play in technical standards?

The EU has already acknowledged its willingness to develop standards on normative issues. In particular, the European Commission (2022d) has recognised that standards no longer only deal with technical components, but also *"incorporate core*

[12] For more information, see Chapter 7.

EU democratic values and interests, as well as green and social principles". Specifically for AI, ethics and fundamental rights questions are at the core of the standardisation strategy. European standards on AI are currently mainly developed by a Joint Technical Committee (JTC) between CEN and CENELEC: JTC 21. But before the creation of a JTC 21, experts on AI within CEN-CENELEC were grouped in a focus group, which notably published a road map, advocating for the creation of a JTC and explaining its vision of European standardisation in the field of AI (CEN-CENELEC, 2020). The road map describes the JTC's future work as *"ensur[ing] that AI is beneficial for citizens and society"*, in particular by developing standards that *"respect fundamental values and human rights"* and *"ensure trustworthy (robust, safe, secure, etc.) AI"*. These objectives have since then evolved into concrete standardisation initiatives[13], and future work will need to address these normative questions.

5.2.4 *Standards are not meant to tackle fundamental rights*

hENs owe their legal existence to Regulation 1025/2012 (European Parliament and Council, 2012) on European standardisation. Regulation 1025/2012 lists the elements that can be considered technical specifications (art. 2.4.a Reg. 1025/2012). The Regulation mentions environmental protection, health and safety, but does not mention ethical criteria or fundamental rights.

The NLF was intended first as a legislative instrument to bring together all the elements of product safety legislation (European Commission, 2022d, p. 12). This emphasis on safety has gradually shifted to include other criteria. The 2022 version of the Blue Guide specifies, in brackets, that *"environmental and health policies also have recourse to a number of these elements"* (European Commission, 2022d, p. 12), but this is clearly a secondary objective of the NLF, which is above all safety-oriented. After "safety" risks, the most commonly addressed risks are health risks, and then, more rarely, environmental risks. Recently, other criteria have begun to appear in the texts on product safety. For instance, Regulation 765/2008 (European Parliament and Council, 2008b) on market surveillance and the marketing of products creates a framework to provide *"a high level of protection of public interests, such as health and safety [...], the protection of consumers, protection of the environment and security"* (art. 1.2 Reg. 765/2008). Regulation 2019/1020 (European Parliament and Council, 2019) on market surveillance and compliance of products, further states that a product should be suspended from free circulation on the market when it presents a *"serious risk to health, safety, the environment or* any other public interest" (art. 26.1(e) Reg. 2019/1020, emphasis added). The term *"any other public interest"* could encompass risks to fundamental rights. However, this is never explicitly stated in the texts.

In Europe, the AI Act is not strictly speaking the first product safety regulation to mention fundamental rights, since the regulation on medical devices also does so. Nevertheless, it is the first time that fundamental rights have been placed on an equal footing with safety and health (Almada and Petit, 2023). In addition, the Medical Devices Regulation was not intended to have standards relating to fundamental rights.

[13] See Section 5.3.1.

5.2.5 *Standards on ethics and fundamental rights are already there*

Although the AI Act is the first product safety regulation under the NLF to highlight fundamental rights, it is not the first time European standardisation organisations have ventured into normative subjects, as standards related to fundamental rights have been used to a lesser extent in the GDPR. Indeed, standards on privacy and data protection, a right enshrined in Article 8 of the Charter of fundamental rights of the European Union, were mandated by the Commission to the ESOs to cover some requirements of the GDPR (Kamara, 2017). However, they were never published in the OJEU, therefore not being granted a presumption of conformity[14].

Initiatives for standards dealing with fundamental rights or the ethics of AI are recent and remain fairly isolated. At present, few standards actively address the social challenges of AI (Mueck et al., 2022). Nevertheless, the trend towards "ethical" standards has developed considerably in recent years, with the first standard explicitly addressing ethics in robotics only dating from 2016 (Bryson and Winfield, 2017). Since then, many stakeholders have sought to develop their own standards on these issues.

However, the analysis of standards on these normative subjects is still virtually absent from the academic literature. In 2017, Bryson and Winfield (2017) examined the first ethical standards at the time, in particular the Institute of Electrical and Electronics Engineers (IEEE) standards. But these initiatives have since multiplied. Other works have listed draft standards for AI, without differentiating between simple technical standards and standards dealing with ethical aspects (Ziegler, 2020a; AI Standards Hub, n.d.). Since the announcement of the AI Act, the European Commission has also published various reports indicating which standards can be used to map to the legal text (Nativi and De Nigris, 2021; Soler Garrido et al., 2023a). In a blog post, Veale (2022) dissects the text of the IA Act to find requirements that could be supported by value-laden standards.

While ethical and fundamental rights standards are increasingly widespread and encouraged by the AI Act, there is a growing need to study the problems of this approach, which we propose to do in this chapter.

5.3 *The plurality of standardisation initiatives on ethics and fundamental rights*

5.3.1 *European standardisation*

European standards for AI today are mainly developed by JTC 21 at CEN-CENELEC. CEN and CENELEC have notably been mandated by the European Commission to develop harmonised standards for the AI Act (European Commission, 2023a). Work on harmonised standards, as well as on other AI standards, is carried out within various thematic working groups. Experts first register with their National Standardisation Body (NSB)[15], before getting involved in European – or international – working groups[16]. The structure of JTC 21 is presented in Figure 5.1.

Despite the apparent desire of the AI Act to extend the scope of technical standards, the European Commission does not expressly refer to a standard on fundamental rights, nor on "trustworthiness" in its standardisation request (SR) for the AI Act (European Commission, 2023a). The standardisation request sets out ten SR items which harmonised standards should cover, corresponding to the require-

[14] See the full story in Chapter 3, Section 3.4.

[15] Such as AFNOR in France (*Agence Française de Normalisation*).

[16] For more information on the overall structure of the standardisation ecosystem and the interactions between standardisation organisations, see Chapter 4, Section 4.5.

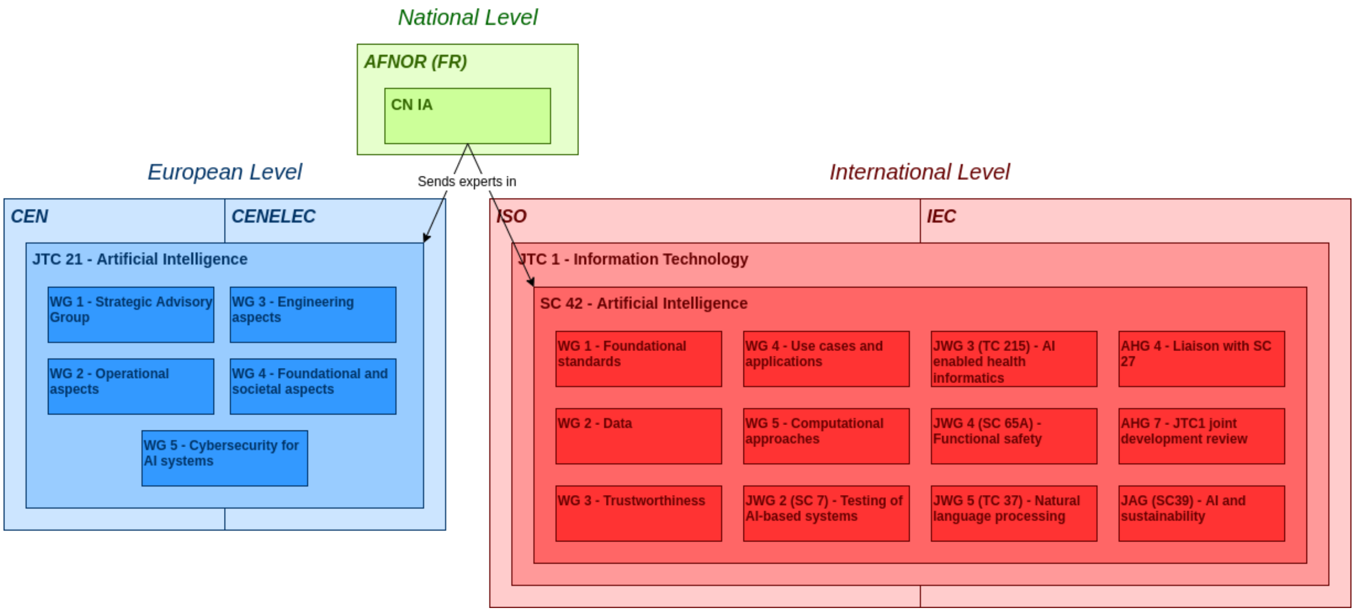

National Level
AFNOR (FR)
CN IA
Sends experts in
European Level
CEN
CENELEC
JTC 21 - Artificial Intelligence
WG 1 - Strategic Advisory Group
WG 3 - Engineering aspects
WG 2 - Operational aspects
WG 4 - Foundational and societal aspects
WG 5 - Cybersecurity for AI systems
International Level
ISO
IEC
JTC 1 - Information Technology
SC 42 - Artificial Intelligence
WG 1 - Foundational standards
WG 4 - Use cases and applications
JWG 3 (TC 215) - AI enabled health informatics
AHG 4 - Liaison with SC 27
WG 2 - Data
WG 5 - Computational approaches
JWG 4 (SC 65A) - Functional safety
AHG 7 - JTC1 joint development review
WG 3 - Trustworthiness
JWG 2 (SC 7) - Testing of AI-based systems
JWG 5 (TC 37) - Natural language processing
JAG (SC39) - AI and sustainability

Figure 5.1: Structural organisation of CEN-CENELEC JTC 21 and ISO/IEC JTC 1/ SC 42. The information used to produce this figure comes from (CEN-CENELEC, n.d.c) and (ISO, n.d.f).

ments for high-risk AI systems set out in Chapter III, Section 2 of the AI Act[17], however, none of them include "trustworthiness". In the request, trustworthiness is rather seen as a cross-cutting theme, not being tackled in a specific standard but being a constitutive part of every standard. This can bee seen, for instance, in the "architecture of standards in response to the EU standardisation request"[18] presented in (CEN-CENELEC, 2024c, p.5), where "trustworthiness" is shown as an horizontal theme, covering all the SR items.

CEN and CENELEC, however, continues to address this topic through a working group within JTC 21, on "foundational and societal aspects of AI systems": CEN-CLC JTC 21/WG 4[19]. The work of WG 4 notably includes a standard called the "AI trustworthiness framework". According to a dashboard released by CEN-CENELEC experts on social media mapping JTC 21 standards to the requirements of the AI Act (CEN-CENELEC, 2024b), the "AI trustworthiness framework" is intended to cover seven of the ten SR items, which makes it very broad and horizontal, a characteristic quite uncommon for technical standards[20]. To date, it is still unclear what the standard will contain, but for the European Trade Union Confederation (ETUC), it is clear that it will have an ethical dimension (ETUC, 2024).

Other works of WG 4 include "AI-enhanced nudging" and "competence requirements for AI ethicists professionals"[21]. Another standard initiative, on "Impact assessment in the context of the EU Fundamental Rights", seems to cover how to conduct FRIA, although the topic is not formally an item requested by the Commission. This shows that ESOs are free to venture beyond the strict limits defined in the Commission's request. Other topics are also been addressed by other working groups. In particular, what were previously ethical principles, such as "transparency", are now also topics of interest. This is because they are also present in the AI Act in the form of obligations for providers of high-risk AI systems. We also note various standards on "data quality", another obligation of high-risk AI systems that was previously part of the "fairness" principle in ethics charters[22]. Standards on quality and risk management also feature prominently in the work of JTC 21. This

[17] For more information on the standardisation request, see Chapter 7, Section 4.7.3.

[18] This architecture of standards was first presented in an Annex to the standardisation request Work Programme of JTC 21 which was not released publicly.

[19] For the structure of JTC 21, see Figure 5.1 or (CEN-CENELEC, n.d.c).

[20] To see experts' opinion on the "AI trustworthiness framework", see Chapter 8, Section 8.5.3.

[21] For a complete list of JTC 21 published standards and standards under development, see (CEN-CENELEC, n.d.b,n).

[22] For a analysis on fairness and data quality in ethics charters, the AI Act and standards, see Chapter 6.

type of standard, which has been very popular since the 1980s[23], is now becoming a trend in standards for AI[24] and software in general (Choi, 2024).

However, CEN and CENELEC are just beginning their campaign to create standards for the AI Act. As of the writing of this thesis[25], a few standards have been published, but these are only ISO/IEC standards adopted by CEN-CENELEC[26] (CEN-CENELEC, n.d.b). Standards fully developed at European level are "under drafting" or "under approval" (CEN-CENELEC, n.d.d). Tensions are starting to arise as international standards are ahead in the development process[27]

Additionally to CEN-CENELEC, ETSI is also working on the creation of standards as part of the AI Act, for example on the evaluation of AI systems, the definition of quality parameters, the explainability and transparency of processing, and the traceability of models (Mueck et al., 2022). However, contrary to CEN-CENELEC which was specifically asked to work on harmonised standards for the AI Act by the Commission, ETSI has been left out of the standardisation request[28]. Although their standards can still be used by companies, they will not benefits from a presumption of conformity.

Standards which are not drafted by the ESOs following a request from the Commission will not become harmonised standards. If this is true for ETSI standards, this is also the case for many international standards and standards emanating from private entities or national institutes. However, these other technical documents may influence the development of hENs for AI by establishing themselves on the market and influencing the state-of-the-art. In addition, standards developed by international standardisation organisations[29] can be adopted by the ESOs and then converted into harmonised standards. Finally, if harmonised standards are not ready in time, the Commission could decide to look elsewhere in the world to implement compliance rules (Baeva et al., 2023)[30]. All these global initiatives are therefore interesting to study because they will have an influence on the European context.

5.3.2 *International standardisation*

At the international level, most of the work on standards for AI is carried out by a sub-committee (SC) inside a JTC shared between the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC)[31]: ISO/IEC JTC 1/ SC 42. ISO and IEC are ahead of the ESOs in developing technical standards for AI: SC 42 has already published 31 standards and 36 more are currently under development[32], counting only horizontal standards[33] (ISO, n.d.f). Within SC 42, standards on normative issues are mainly developed by WG 3 on "Trustworthiness"[34], an equivalent to JTC 21 WG 4. Other important standards are also being developed by SC 42 WG 1 on "Foundational standards", which define the main terms and concepts and will serve as a basis for other standards.

These SC 42 standards for AI include the same types of standards as in JTC 21. First, many initiatives address broad normative issues – such as ISO/IEC 24368 (2022d) and ISO/IEC 22443 (n.d.b) on the social and ethical concerns regarding AI, or attempt to define concepts around "trustworthiness" such as ISO/IEC 24028 (2020). Other standards take a more practical approach, looking at different types of biases present in AI systems – such as ISO/IEC 24027 (2021), defining methods to mitigate them – such as ISO/IEC 12791 (n.d.e), data quality criteria – ISO/IEC 5259-1 (2024b) and others in the 5259 series, or interpretability methods – such

[23] See Chapter 4, Section 4.4.1.

[24] For a discussion on this type of standards and their shortcomings, see Chapter 9, Section 9.3.

[25] November 2024.

[26] CEN and CENELEC can adopt ISO and IEC thanks to international agreements. This process is explained in Chapter 4, Section 4.5.1 and Chapter 7, Section 7.5.2.

[27] For more information, see Chapter 7, Section 7.6.2.

[28] For more information on the reasons of this absence, see Chapter 7, Section 7.9.3.

[29] Such as ISO and IEC, see next paragraph.

[30] This possibility could happen in the context of "common specifications". For more information on common specifications, see Chapter 7, Section 7.9.4.

[31] For more information on ISO and IEC, see Chapter 4, Section 4.5.1.

[32] As of the writing of this thesis in November 2024.

[33] ISO/IEC JTC 1 on "Information technology" has several subcommittees, such as SC 42, which works on standards that apply to all AI systems. But there are other subcommittees dedicated to certain AI applications, such as SC 37 on biometric systems (ISO, n.d.e), or SC 29 on audio, picture and multimedia applications (ISO, n.d.d). Here, we are simply counting the work of SC 42.

[34] For the structure of SC 42, see Figure 5.1 or ISO website (ISO, n.d.f).

as ISO/IEC 6254 (n.d.c), inspired in particular by the ethical principles of fairness and explainability. Finally, some standards adapt risk management to AI – such as ISO/IEC 23894 (2023a) and ISO/IEC 42001 (2023b), while some others adapt IT governance framework to AI – such as ISO/IEC 38507 (2022c). Outside of SC 42, ISO also develops standards on broad ethical and social aspects which can apply in the context of AI, such as ISO 26000 (2010) on "social responsibility".

Apart from ISO and IEC, the International Telecommunication Union (ITU) is also working on AI standards, but less directly linked to ethical aspects, such as AI for cloud computing, quantum computing or network applications (ITU, 2022). ITU is also co-leading the United Nations (UN) Inter-Agency Working Group on AI (IAWG-AI) (UN CEB, n.d.). The IAWG-AI will for instance be involved in a newly established UN Task Force in charge of developing *"guidance that aligns with existing UN principles and standards on ethical AI use"* (UN CEB, 2024).

5.3.3 *National institutes*

Outside the well-defined framework of standardisation bodies, some entities are developing their own initiatives to standardise or certify the ethics of AI. This is the case of the national institutes, which are working on the development of evaluation frameworks for AI, independently of the European and international standardisation bodies.

In the United States, the National Institute of Standards and Technology (NIST) is also working on AI standards (NIST, 2023b). One of its major projects is the Face Recognition Vendor Test[35], which lays the foundations for evaluating facial recognition systems. The NIST compares the performance of dozens of algorithms from different manufacturers around the world[36]. These comparisons are based on different accuracy criteria and, in terms of algorithmic fairness, on mathematical measures of demographic differences (Grother et al., 2019; Grother, 2022). Metrics such as the divergence rate or the ratio of error cases between two populations can be used to calculate the severity of a bias. The tests carried out by the NIST are closely followed by companies and are a decisive selling point for customers. The measures they select to evaluate AI systems therefore have a good chance of becoming the reference standard in the field. In particular, ISO could draw on them in the context of its subcommittee 37 on biometrics and its standard on quantifying the variation in performance of biometric systems in demographic groups (ISO/IEC, 2024a). In 2023, the NIST published a risk management framework to *"help align the decisions about AI system design, development, and uses with intended aim and values"*. In particular, it outlines the *"characteristics of trustworthy AI systems"*, as well as actions to ensure that they are put into practice (NIST, 2023a). More recently, a new version was published, specifically for generative AI (NIST, 2024). In addition to their widespread adoption by the market, it is likely that NIST publications will continue to grow in importance as they are increasingly cited in official communications from US government departments[37].

But the NIST, although the best known, is not the only national institute in the world to develop AI standards. In France, the Laboratoire National de Métrologie et d'Essais (LNE) has developed its own "Certification Standard of Processes for AI" (LNE, 2021). It defines a number of requirements to be met during the design, development, evaluation and maintenance in operational conditions of AI systems. The standard places particular emphasis on the transparency of processes, listing

[35] All FRVT are available on the NIST website (NIST, n.d.b).

[36] The results are available online, and a visualisation tool shows which company is in the lead (NIST, n.d.a).

[37] Such as for instance the Department of State (Bureau of Cyberspace and Digital Policy, 2024).

the elements that must be documented and communicated to the customer. Furthermore, it is not based on a particular technology but on a way of operating at company level. While there is no particular desire to deal with ethics, it nevertheless reflects the desire to establish a set of best practices within AI companies. It also enables a certificate to be issued attesting to the fact that the company meets the conditions set out by the standard. Some French companies have already received this certificate[38].

5.3.4 *Professionals associations*

In addition to these national institutes, a number of professional associations are developing their own guidelines. Among the international initiatives on AI ethics, the most important is undoubtedly that of the Institute of Electrical and Electronics Engineers (IEEE), an organisation engineers specialised in information and communications technology (ICT). Although IEEE is a global organisation, it is based in the US and its standards are typically recognised by the American National Standardisation Institute (ANSI) (Gonzalez Torres and Ali-Vehmas, 2024). In 2019, the IEEE published a document containing a list of ethical principles and building a framework to make them operational (Chatila and Havens, 2019). In particular, the report presents recommendations and guidelines with the intended purpose to *"inspire the creation of standards [...] and associated certification programs"*. The report was the first step in the IEEE Global Initiative (IEEE, n.d.c) on the ethics of autonomous and intelligent systems and was indeed followed by a massive development of standards relating to the ethics of AI systems. The first to be published in 2021 was the IEEE Std 7000™-2021 (IEEE, 2021a) standard, on ethical concerns during system design. Other standards would follow, more specialised on an ethical aspect, such as the IEEE 7001™-2021 standard on transparency (IEEE, 2021b). Some are still in development, such as the IEEE P7003™ standard on algorithmic biases (Koene et al., 2018). In total, the project plans to develop more than fifteen standards relating to the ethics of AI[39]. In addition to these "ethical" standards, the IEEE is also developing a certification program for AI ethics, called CertifAIEd (IEEE, n.d.b), based on the determination of a risk profile for the AI system, followed by an evaluation according to a series of criteria.

Other initiatives are also worth noting, such as the Verband der Elektrotechnik (VDE) in Germany, which has developed its own framework for "AI trustworthiness characterisation" (Peylo et al., 2022). It takes values such as transparency, accountability, privacy, fairness and reliability, and assigns measurable criteria and indicators to them. A score can then be calculated, representing the level of trust in the system.

It has become frequent that part of the companies contributing to traditional standardisation initiatives, decide to form a consortium to supplement the formal standard setting process (Baron et al., 2014). They either develop their own standards or accompany official SDOs. This allow to align companies positions and improve their chances to influence standards. Among these groups, the World Wide Web Consortium (W3C) as already position themselves in favor of standards at the intersection of AI and the web (W3C, 2024). In an online document, they build on ethical principles to define the need for AI standards for the web, including aspects such as autonomy, transparency, privacy, safety and so on. W3C's community group on "Artificial Intelligence Knowledge Representation" has notably examined

[38] This is for instance the case of Axionable (2021).

[39] These standards are cited on IEEE website (IEEE, n.d.a).

the possibility of a standard for explainable AI (Di Maio, 2018), but no particular initiative seems to have emerged as of now from the consortium. Another consortium, the Internet Engineering Task Force (IETF), has also taken an interest in AI, notably with its latest initiative on AI for network management (Martinez-Julia et al., 2021). Given that human rights advocates have already collaborated with the IETF on some of their work (Cath-Speth, 2021), it is not impossible that a similar trend will emerge with AI.

5.3.5 *Private companies*

Some companies are also developing their own evaluation frameworks. For example, Microsoft published last year its Responsible AI standard (Microsoft, 2022) which lists various requirements relating to accountability, transparency, fairness, reliability, privacy and inclusiveness. Broader frameworks are sometimes published in the form of corporate AI policies[40]. However, these are often simply recommendation documents and not real evaluation benchmarks. The plurality of such documents makes it hard to differentiate between simple ethics charters and real technical standards.

Other smaller companies are also developing their own evaluation frameworks for AI systems, setting up ethical AI labels, based on questionnaires or algorithmic audits. In France, for instance, GoodAlgo proposes a label called ADEL, to assess the ethics of AI systems based on compliance with a certain number of criteria (GoodAlgo, n.d.). Some labels focus on more specific criteria, such as the "human guarantee"[41] label for AI in health, supported by the Digital Medical Hub and the company Ethik-IA (E-Santé, 2023), or the GEEIS IA label for equal opportunities (Arborus, n.d.).

These labels are easier and quicker to set up than certification processes based on official standards, and allow consumers to make an informed choice while waiting for an official certification. However, the criteria for assessing systems and issuing labels are often not public. The diversity of approaches can help to democratise AI ethics, but it can also lead to a collision of different visions. By analysing these different approaches, we can detect dynamics and visualise the different directions that AI ethics can take: an approach based on corporate governance and management, putting in place best practices for the planning, design or monitoring of AI systems, risk management and anticipation of social challenges; or an approach based on the measurement and evaluation of the technical performance of AI systems.

5.4 *Problems with standards on ethics and fundamental rights*

5.4.1 *It is hard to separate a technical question and an ethical or a fundamental rights question*

It is not always easy to assess the ethical nature of a standard or to see when technical questions may have an impact on fundamental rights. ANEC[42] the organisation that defends the interests of European consumers in standardisation matters, has already recognised the many difficulties involved in transposing EU fundamental rights and values into technical standards (Giovannini, 2021). In an ideal world, technical standards should be separated from *"hard normative questions"* (Laux et al., 2024) and value judgements. In reality, however, it is hard to separate the two.

[40] See for instance SAP (2024) policy on AI. Many more of these documents are listed in our database of AI ethics documents: `http://mapaie.telecom-paris.fr`.

[41] Translation from French by the author.

[42] As stated on ANEC's website: *"ANEC stands for the 'European Association for the Co-ordination of Consumer Representation in Standardisation AISBL' [...] ANEC is often described as 'The European consumer voice in standardisation"'* (ANEC, n.d.).

For instance, the concept of fairness in AI systems has several meanings, both morally, legally, and technically (Mulligan et al., 2019). In a general sense, fairness means *"the quality of treating people equally or in a way that is right or reasonable"* (Cambridge Dictionary, 2024). This relates in law to the principle of non-discrimination protected by Article 21 of the EU Charter of Fundamental Rights (European Commission, 2012a) and Article 10 of the Treaty on the Functioning of the European Union (European Commission, 2012c). There are many technical definitions of fairness, and a system that is fair according to one definition is not necessarily fair according to another. Many definitions cannot even be satisfied at the same time (Chouldechova, 2017b). For instance, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, used in the United States to predict the recidivism rate of criminals, has been accused of penalising African-Americans according to a certain fairness criterion (Angwin et al., 2016), whereas it respected fairness according to another measurement method (Northpointe, 2019). By defining technical formulas to measure fairness in a standard, we run the risk of choosing an approach to non-discrimination that will lead to injustice in certain situations. This example shows that a seemingly technical definition of fairness can hide a normative choice affecting fundamental rights[43], the kind of normative choice that generally is made by lawmakers and judges.

Another example is the NIST study on demographic differential for facial recognition (Grother, 2022) which displays a few "equity measures" for facial recognition systems. For all of them, error rates are calculated for different groups of people, based on sensitive personal information like gender or ethnicity. For example, some measures are based on a comparison between the error rates of the two groups on which the system performs best and worst, and other measures are based on the average of all error rates. The first case is, unfortunately, not very robust, and even a slight change in parameters can produce a totally different result. On the contrary, an average-based measure will be more robust but will erase the difference between groups: a system whose performance is very poor in one group but excellent in the others, could end up with the same score as a system whose performance is correct in all groups. Thus, the poor performance of this one group could go unnoticed. Yet, if a system does not work well for a certain category of population, it can lead to discrimination, such as people of colour being wrongly accused of committing crimes because an algorithm has matched their face to that of a criminal (Hill, 2020b).

The NIST (n.d.a) also proposes a benchmark that evaluates the fairness of systems against their performance. A manufacturer can choose to focus on optimising their score in the given performance or fairness criteria. They can also choose which fairness metric they should improve: the benchmark includes demographic variations by false match rate (FMR) or false non-match rate (FNMR). A low FMR aims to avoid mistakes where a person is wrongly judged to be the same as in a certain image, which usually involves higher security and social stakes to avoid intrusions into a building or station, and false accusations in case of police use. A low FNMR avoids systemic rejection of certain people.

A choice of standard signals a preference for a specific logic and set of priorities (Timmermans and Epstein, 2010). Standards organise social life, and it is crucial to question what choices have been made and how they could have been made differently (Timmermans and Epstein, 2010)[44]. However, in the choices made in AI standards are often presented as purely technical, and therefore non value-laden,

[43] Benbouzid (2023) explains that fairness approaches seek objectivity, based on the justification from the system designer as to the "reasonable" level of bias, but that their results are in fact politically situated.

[44] Since even the most technical standards have social implications, according to Winfield (2019b), all standards can therefore be considered as implicit ethical standards.

choices (Solow-Niederman, 2024). By trying to define good ethical behaviour in technical standards, we risk reducing ethics to a set of tools, which trivialises moral reasoning (Bietti, 2020; Balayn et al., 2023)[45].

5.4.2 *Compliance to standards can lead to ethics washing*

The diversity of approaches to AI ethical development, such as the multitude of fairness measures, is likely to lead to strategic simplification choices (Aïvodji et al., 2019). Manufacturers will display the measure that shows that their system is free of bias and therefore fair according to them and not the other measures showing the system is discriminatory[46]. The introduction of mathematical measures of fairness in a standard, for instance, is likely to accentuate this trend, by giving greater legitimacy to any chosen measure included in the standard. Furthermore, many entities tend to adopt standards symbolically, without changing their practices (Cihon et al., 2021). Compliance with these standards then becomes a simple sales argument, and the recent development of "ethical" certifications and labels legitimises these practices and perpetuates a culture of ethics washing (Wagner, 2018). Additionally, the protection granted by standards is limited and having in place a risk management system will not guarantee that all possible harms have been taken into account, or that the protective measures are sufficient. For instance, respecting a mathematical notion of fairness does not guarantee that the system will not discriminate (Hoffmann, 2019). Some people therefore fear that the standards developed at European level for the AI Act will not be are not restrictive enough, leaving the industry to do as it pleases (Pouget, 2023).

Certification to technical standards is often perceived by consumers as a guarantee of safety (de Tervueren, 2012). This is particularly true of CE marking, often regarded as the cornerstone of the European trustworthiness model, a system that European citizens have come to internalise and respect (Burden and Stenberg, 2022). But the mark is also often wrongly understood by consumers as a guarantee of quality when in fact it only signifies compliance with regulations. Indeed, studies have shown that it is difficult for citizens to understand what the CE mark represents (Burden and Stenberg, 2022). Products covered by the NLF do not require pre-market approval to be sold in the EU. The CE mark therefore does not indicate that a product has been approved by a government agency or by the EU (European Commission, n.d.c). As recalled by the Blue Guide, CE marking is a key indicator of a product's compliance with EU legislation, but it is not a proof of that compliance (European Commission, 2022d, p.64). As such, a CE marked product may also have safety flaws. Several high-profile cases have involved medical devices – breast implants (van Leeuwen, 2014; Rott, 2019) and glucose monitors (Wentholt et al., 2005) – that had the CE marking but which were seriously defective[47]. In the same way as for CE marking and safety standards, it is likely that a CE marking relating to fundamental rights may be incorrectly interpreted by citizens as meaning that a given AI system respects fundamental rights.

However, complying with a mathematical notion of fairness as defined in a standard does not guarantee that the system will not discriminate. Similarly, complying with a standard which includes an obligation to conduct an impact assessment makes it possible to anticipate certain consequences of deploying a system, but this does not mean that all possible prejudices have been taken into account, or that the protection measures taken are sufficient. The certification mark could therefore give system users and citizens a false sense of protection.

[45] See also our discussion in Chapter 9, Section 9.2.1.

[46] This practice is know as "ethics shopping" (Wagner, 2018).

[47] For more information on these cases, see Chapter 3, Section 3.6.4.

5.4.3 *Cultural and contextual aspects of ethics and rights constitute an obstacle to the development of general standards*

The EU is based, from a constitutional standpoint, on core democratic and economic values listed in Article 2 of the Treaty of the European Union (TEU) (European Commission, 2012b), which includes *"the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights"*. The EU's identity was built on the definition of this set of values which, in the case of digital technology, combine with technical criteria which make their application more complex (Keller and Levallois-Barth, 2021). Furthermore, these values are not necessarily shared among international actors. For instance, Gonzalez Torres and Ali-Vehmas (2024) compares the values displayed by public authorities and governments in the US, China and the EU, and reveals major differences. Even when similar values are used, they may have different meaning in different parts of the world (Fung and Etienne, 2022). As a result, the technical methods in which these values are embedded do not necessarily comply with European norms. For instance, most of the tests designed to detect biases in AI do not live up to the standards of European non discrimination law, because these tests were developed in the context of US legislation (Wachter, 2024). ANEC therefore calls on the Commission to develop its own standards in order to *"safeguard European values or ethics"* (Russell, 2021). It also stresses the importance of not jeopardising fundamental European values for the sole purpose of reducing development times. This means not giving international bodies sovereignty over standards and taking the time to develop European standards (Cihon et al., 2021)[48].

Moreover, the development of these standards on ethics and normative questions is not universally accepted. According to Tessier (2022), the adjective "ethical" refers to morality and can therefore only be applied to *"an approach, a deliberation, a reflection, a question, a principle, a value"*[49]. So the notion of "ethical conformity" or "ethical standard" is questionable, because ethics is inherently variable. Fairness, for instance, is too contextual for requirements to be defined in general (Wachter et al., 2021b; Weerts et al., 2023). Moreover, AI ethics is dynamic and a system cannot be perpetually aligned with ethical values (Bryson, 2022). In this respect, static certification of AI ethics cannot capture any relevant issue. An AI certified as "ethical" will remain so even if the system and its context evolve. Although some organisations may wish to regularly re-evaluate the system and keep a certificate up to date[50], such a mark can never be fully adapted to a specific context, otherwise it would lose its universally applicable character.

Similarly for fundamental rights, Smuha and Yeung (2024) explains that, unlike safety risks which can be *"materially observed and measured"*, fundamental rights are *"political constructs"*. Their legal protection is therefore subject to an evaluation of the alleged interference or infringement, in which the specific context is scrutinised. Ultimately, standards on ethics and fundamental rights are bound to fail if they remain too general and do not take into account the cultural, contextual, sector-specific and systemic aspects of the technology there address.

[48] For more information on the influence of international bodies on European standardisation, which is reinforced by the time constraints of AI standards, see Chapter 7, Section 7.5 and Section 7.6.2.

[49] Translated from French by the author.

[50] This is the case of the LNE framework for instance (LNE, 2021).

5.5 Can AI standards address ethics and fundamental rights (and how)?

5.5.1 AI standards can fall into three different categories

Our overview of standardisation initiatives reveals different types of standards. From this analysis, we draw our own classification of standard types, represented in Table 5.1.

Governance standards	Information standards			Performance standards
	Ethical frameworks	Trustworthiness frameworks	Technical reports	
On: processes	On: ethical concerns	On: products and processes	On: products	On: products
Include: quality management, risk management, conformity assessment...	Include: potential harms, ethical stakes references, general recommendations...	Include: list of properties, definitions, general recommendations...	Include: definitions, metrics, methods, benchmarks...	Include: minimum threshold, choice of benchmark, choice of metric...
Ex: ISO/IEC 42001:2023b	Ex: ISO/IEC 24368:2022d	Ex: CEN-CENELEC AI trustworthiness framework	Ex: ISO/IEC 24027:2021	Ex: None

Table 5.1: Classification of standard types

The first type is what we will call "governance standards", which deal with the mechanisms to be put in place at an organisational level to ensure that ethical principles are effectively respected. They are often more about the processes of AI development than about the AI product itself. These standards are often an adaptation of quality management standards that were trending in the 1980s[51]. An example of such a standard are the ISO/IEC standards such as 42001 (2023b), or some European standards such as the one on conformity assessment. But governance standards also include for instance some of IEEE standards, such as IEEE Std 7000™-2021 (IEEE, 2021a), which takes into account the consideration of the different stakeholders and their values during the exploration and development phases of systems.

The second category of standards corresponds to what we call "information standards", which give descriptive elements without giving prescriptive quantified choices. But the information provided in these standards can have different levels of technicality. Some standards remain very broad, not even incorporating technical elements, but merely describing social or ethical stakes, such as in ISO/IEC 24368 (2022d) and 22443 (n.d.b). Others go a little further than stating the problems and concerns and look at what the properties of a "good" AI might be. This is the case, for instance, with standards on trustworthiness, such as ISO/IEC 24028 (2020) or the AI trustworthiness framework of CEN-CENELEC. However, these standards seem to stay at a very high level, listing properties that AI systems should meet but without giving much explanation about their implementation. For instance, ISO/IEC 24028 (2020) discusses "transparency", "explainability", "reducing biases", "privacy" or "functional safety" as part of the "mitigation measures" to "vulnerabilities, threats and challenges" previously identified. Similarly, it is

[51] For more information about the evolution of standardisation through the ages, see Chapter 4, Section 4.4.1.

likely that the AI trustworthiness framework from CEN-CENELEC will address each item of the standardisation request, such as "risk management", "quality of datasets", "transparency" or "human oversight" and so on. However, there is a risk that these frameworks remain very broad and do not go into much more technical detail than ethics charters. Finally, information standards can also give an overview of a domain, with relevant technical definitions, metrics or methods. For example, ISO/IEC TR 24027 (2021) provides numerous definitions of terms related to bias and fairness, as well as mathematical formula to assess bias and finally list a few methods for the treatment of "unwanted" bias. Although it does not take a position on the best measure to adopt, this standard inevitably includes an ethical aspect because of the subject it deals with. It is a "technical report" (TR)[52] and as such, it resembles a technical state of the art. However, it should be noted that these standards are not always considered as such by SDOs, which refer to them instead as "reports" or "specifications"[53]. Additionally, although we stated that these standards only describe but do not prescribe, it is not entirely true. SDOs distinguish between "requirements" and "guidances" depending on the degree of importance of the prescriptive element[54]. Both are found in information standards. However, there is no *quantified* prescriptive elements, as a requirement can just be that companies "shall choose the best metric in their context".

In both cases, AI standards do not dictate the results expected from an AI system: for the moment, there is no such thing as a "performance standard"[55]. Current standards highlight good practice, whether in the process management or in the technical methods of design and assessment. This way, they address ethics and fundamental rights but do not directly answer the hard normative question of what should be deemed acceptable.

5.5.2 *AI standards cover ethics and fundamental rights through good practices*

ANEC has already advised that hENs should not be used to define or apply fundamental rights, legal, or ethical principles (Giovannini, 2021). However, if standards cannot directly attest to respect for fundamental rights, what purpose do they serve and what should they contain?

Let us take the example of a standard on fairness. Such a standard can be used by a company to benchmark itself against the competition and assess its own progress. If the results are good enough, the company will use the standard as a marketing tool, like the NIST benchmark for facial recognition for which companies compete to achieve the best results based on different fairness tests. This fosters competition between companies and encourages them to innovate (Blind, 2016). A standard can also enhance transparency and redress information asymmetries (Cantero Gamito, 2018) by presenting to users and citizens a standardised score of different performance parameters, including for fairness, thereby permitting better comparison between products. Finally, standards, such as hENs, that are linked to legal compliance obligations, provide public authorities with a uniform method for assessing compliance.

These different uses of standards hint to what they can and cannot contain. For compliance, hENs will help clarify the AI Act's approach to risk, for instance by defining how to conduct a risk management system, or detail what elements a conformity assessment should contain[56]. Additionally, standards can help harmonise how to conduct an algorithmic impact assessment (Calvi and Kotzinos, 2023) or a

[52] See the different types of deliverables in Chapter 4, Section 4.5.3.

[53] This is because we took a broader definition of the term "standard" that the SDOs' definition. For more information on the reasons of this choice, see Chapter 4, Section 4.6.2.

[54] See Chapter 4, Section 4.6.2.

[55] Following Allen and Sriram (2000) terminology, also referred to as quality standards Blind (2004).

[56] These topics are notably present in the standardisation request (European Commission, 2023a).

FRIA. As regard to governance, standards can provide guidance on the structure to be put in place within the company – perhaps with a digital ethics officer or an ethics board, the competences required for this position, or the type of decisions they can and cannot make[57]. Product-based standards can define tools to help make better design decisions. For example, they can define all the evaluation measures known in the literature[58] – paying attention to selection biases, or the technical means to avoid a system malfunction that could lead to fundamental rights violations in the long term. In short, standards can help define tools and provide a common vocabulary for comparison between products or companies. These tools can help market actors transparently compete on fundamental rights issues, showing they have responsible processes in place, and that on certain metrics, they have achieved a certain score on an issue such as fairness. Laux et al. (2024) similarly propose that standards provide for *"ethical disclosure by default"*, a system guaranteeing that users, regulators, judges, and other stakeholders receive meaningful information in order to evaluate fundamental rights compliance in a given context.

5.5.3 *AI standards should not try to set thresholds*

However, there are some things that AI standards should not try to do. Even when following a standard on risk management, the evaluation of risks will remain under the responsibility of the provider. A standard can therefore never say what risks are acceptable or unacceptable (Fraser and Bello y Villarino, 2023). Fairness standards, should not say what definition of fairness should be used for a given use case[59] or what the acceptable threshold of unfairness is. In case there is a trade-off to be made between fairness and performance, a standard should not say what that trade-off should be. A standard can only provide different ways of defining and measuring fairness, making sure everyone is using the same taxonomy and methodology to measure the different aspects of fairness, but will not say which aspect of fairness should be given priority, or whether a residual level of unfairness can be tolerated in a given situation.

Performance standards are quite common in product safety. They specify how the product is to be built, what materials are to be used, how they are to be assembled, and so on. They also specify the tests the product must meet, such as the exact temperature or pressure it must withstand. In product safety, it is not unusual for a standard to define a threshold, for example a level of resistance to fire, or the error rate of a safety component for machinery. These standards are, however, nearly impossible to establish today for AI systems due to their probabilistic nature, which makes their reaction to certain tests highly dependent on the situation, the data on which the system has been trained, etc. This is even truer for standards that have a direct impact on people's fundamental rights, such as fairness standards. Setting a threshold for these measures would be like setting a threshold for the level of discrimination that may be accepted: it is neither a universal decision nor something acceptable from a legal standpoint. Setting a fairness threshold could also be abused by claiming that a system is "fair enough", without any concern for improving fairness further (Buyl and De Bie, 2024).

Using technical standards to ensure respect for fundamental rights is therefore very risky, because the standards could compete with the work of the legislator and judges, who are the only ones competent to assess and balance interference with different fundamental rights (Gornet and Maxwell, 2022). A technical standard designed to protect personal safety is also intended to preserve a fundamental right,

[57] Such as, for instance, the standard on "competence requirements for AI ethicists professionals" that is being prepared by CEN-CENELEC.

[58] Such as the upcoming ISO/IEC standard on bias mitigation ISO/IEC (n.d.e), which will likely list all the means known in the literature for assessing and dealing with bias.

[59] This includes both the metric used and the population groups on which the system is evaluated.

in particular the right to life. But a technical standard on fire resistance, for example, will conflict less with the primary role of judges. Of course, a judge will always be able to consider that the application of a safety standard in a specific case was insufficient. But the technical standard on safety will not be perceived by judges as interference in their work, whereas a standard on respect for fundamental rights might be. As a result, whether a fairness score is acceptable or is the right metric to be using in this situation should remain outside of standards and determined by the regulator and judge.

As thresholds cannot be set for standards relating to the protection of fundamental rights, the development of hENs on these subjects for the purpose of assessing compliance with the AI Act seems like a difficult – and not necessarily desirable – task. Because of their legal effects, hENs will always aim to set thresholds, and that indeed seems to be the intent of the AI Act since hENs and CE marking are supposed to signal compliance (Laux et al., 2024). But outside of the safety realm, hENs are less suitable, as they cannot define what is an "acceptable" level of protection to fundamental rights. Standards should not attempt to answer these hard normative questions, nor should they seek consensus; they should rather create means of disclosure (Laux et al., 2024). Access to information regarding a certain technology can then enable regulators and judges to make specific decisions in a given context. We therefore invite standardisation actors to develop standards, whether hENs or other standards, which contribute to the protection of fundamental rights through the dissemination of good practices, but which avoid making strong value-laden societal judgements.

5.6 Conclusion

This chapter shows the AI Act's attempt to operate at two levels: ex ante compliance, inspired by product safety rules with the use of hENs and CE marking, and the protection of fundamental rights through "trustworthiness", i.e. ethical and fundamental rights properties translated into technical criteria. However, product safety tools such as hENs and CE marking are not meant to cover fundamental rights, nor ethical properties of technical systems. And indeed, standards have difficulty in addressing ethical and fundamental rights issues, and when they attempt to do so they can lead to ethics washing and consumer deception. Nonetheless, as AI raises multiple ethical, social and legal issues, standards will need to address these normative questions. Despite these difficulties, this does not mean that standards cannot address fundamental rights, as they still have an important role to play in encouraging best practices in processes and measurement techniques, but they can never attempt to decide on a trade-off or on a level of acceptability of a given fundamental right risk.

The AI Act approach calls into question the very nature of standards and their limits. It might also pose problems for the interpretation of standards by the courts, as in the past the boundaries between the technical and legal worlds were well-defined, whereas today there is a certain overlap. In this context, even more than in the case of safety standards, ESOs will have to account for the power they hold. The hENs to be developed in support of the AI Act will set the tone for future regulations in the field of digital law. Europe should, however, be cautious about the power it grants to hENs, particularly if they continue their foray into fundamental rights.

Chapter 6
Case study: fairness standards, the new ethics charters

Where does this chapter come from?

This chapter is an original contribution to the thesis. It aims to illustrate the difficulty to standardising ethics, by focusing on one of its core principle: fairness, and look at how this principle is tackled by technical standards. We choose to focus on fairness because of its central status in ethical AI discussions and, if not achieved, AI systems can have harmful consequences for individuals[1]. It should therefore be at the center stage of AI regulation. The data and information collected for this work on standards was last updated in March 2025.

Chapter's abstract

After a wave of ethical charters in recent years to define the values and principles to be respected in the lifecycle of an artificial intelligence (AI) system, most studies agree that fairness is an essential component to make AI more "trustworthy". In addition to these contributions, the legal field has also taken an interest in fairness. In Europe, the recently adopted legal framework for AI – the AI Act, requires providers of "high-risk" AI systems to demonstrate that their AI is free from biases. Under the AI Act, technical standards which seek to harmonise definitions and practices will play a crucial role in assessing compliance.

In this chapter, we provide an overview of standardisation initiatives which address fairness for AI systems, and analyse their content and progress. We show that while progress is being made to bring standards in line with ethical and legal expectations, the work of AI standardisation – and especially AI fairness standardisation – is still in its infancy. Finally, from ethics to law and standards, fairness has lost some of its meaning and scope. The transition from one regulatory mechanism to another is therefore not made without a cost.

[1] See Chapter 1, Section 1.1.6.

6.1 Introduction

Recent years have seen the emergence of charters and manifestos on AI ethics, containing principles and guidelines to be applied throughout the lifecycle of an AI system. The late 2010s have notably been very prolific in ethics charters. In 2019, Jobin et al. analysed 84 charters; in 2022, the Council of Europe would list over 600 of them on their website[2]. The theme encompassing justice, fairness and non-discrimination, equity, diversity and inclusion, is among the main ethics principles almost always mentioned in these charters (Jobin et al., 2019; Fjeld et al., 2020).

While in computer science the field of algorithmic fairness, mainly understood as the detection and mitigation of undesirable biases in machine learning models[3], has grown exponentially these past few years (Zhang et al., 2022), the social sciences have been exploring this concept for decades with a much broader meaning, defining it as a matter of societal organisation, moral duties or positive legal rules, depending on the discipline (Mulligan et al., 2019). The recent appropriation of the concept of fairness by computer scientists, and the need to cross disciplines to get the full picture, have led to debates about the true meaning of the term, the compatibility of fairness measures with law (Weerts et al., 2023; Wachter et al., 2021a,b), or their comparison with philosophical frameworks (Binns, 2018). This interdisciplinary emulsion has been particularly evident at new scholarly conferences and events, focusing, among other things, on ethics and fairness, such as the ACM[4] Conference on Fairness, Accountability, and Transparency (FAccT)[5] or the AAAI[6]/ACM AI, Ethics and Society conference (AIES)[7].

In summer 2024, the regulation of AI took a step further in Europe with the adoption of the first mandatory framework for these technologies: the AI Act (European Parliament and Council, 2024b). The AI Act draws heavily on ethics charters and policy documents and proposes to regulate AI through risk. This risk-based approach will ensure that systems with different risks to safety and fundamental rights comply with various levels of requirements. For instance, "unacceptable risk" systems are totally prohibited, while low-risk systems are only encouraged to comply with voluntary codes of conduct. Compliance is, however, more complex for two categories of systems: "high-risk" systems and "General Purpose AI" (GPAI) models, which will have to comply with a number of technical requirements[8]. Among the obligations laid down by the AI Act for high-risk systems, one is closely linked to the principle of fairness: the requirement on data governance (Art. 10).

Standardisation organisations, responsible for developing technical standards, have also recently begun to take an interest in AI. A number of working groups have been set up within international bodies, as well as at European and national level. These initiatives, particularly in Europe, have multiplied since the publication of the first version of the AI Act in 2021 (European Commission, 2021e). Indeed, some of these standards are intended to apply directly alongside the AI Act, giving them an additional legislative function[9]. Standards will therefore play a crucial role in the years to come, as they are heavily used by companies and will partly determine the technology, development and management practices that the industry will follow.

Among the many themes addressed by these standards, fairness is not the most common, unlike the well-discussed topics of data management, interoperability and system architecture[10]. This leaves a gap between the guidelines in ethics charters,

[2] We have also analysed these AI ethics charters in Chapter 2.

[3] Machine learning is a family of algorithmic methods and a subdivision of artificial intelligence in which the model learns by training on data. For more information, see Chapter 1, Section 1.1.3.

[4] Association for Computing Machinery

[5] `https://facctconference.org/`

[6] Association for the Advancement of Artificial Intelligence

[7] `https://www.aies-conference.com/`

[8] For more information on the categories of AI systems under the AI Act, see Chapter 3, Section 3.5.2.

[9] See Chapter 4, Section 4.6.2, or our article (Gornet and Maxwell, 2024).

[10] These three topics are the most frequently used labels in the AI Standards Hub (n.d.) database.

the legal requirements of the AI Act, and the standards that companies will follow in the development of AI systems.

In this chapter, we do an overview of existing fairness standards, their content and progress, to see to what extent they cover all the realities that the term "fairness" encompasses.

We start in Section 6.2 by examining related works and literature. We show that few works have addressed the issue of AI standards and, when they do, it is often in a general way and not specific to AI fairness. In Section 6.3, we describe our methodology for this study, including how we selected fairness standards for analysis, and show that ethics charters often separate the principle of fairness into three broad categories: Bias, Accessibility and Participation. We then present, in Section 6.4 and Section 6.5, the results of two complementary analyses: a quantitative analysis based on a list of ethics and fairness standards that we collected and annotated, and a qualitative reading of a few selected standards. The first enables us to analyse the progress made, and to verify that the theme of fairness has been addressed in greater depth over the years. The second enables us to delve deeper into the content of these standards, to identify and compare trends. Finally, in Section 6.6, we examine two tendencies of current standards: (1) the tendency to address many topics at once, including fairness and other ethics principles, which resembles AI ethics charters, (2) the tendency to focus on specific aspects of fairness such as data quality and bias mitigation. We present the limitations of this study and avenues for future work in Section 6.7, and conclude by summarising major findings in Section 6.8.

6.2 Related works

6.2.1 Fairness in ethics charters and in legal frameworks

Many studies have examined ethics charters for AI, either creating databases (Council of Europe, 2022; Algorithm Watch, n.d.; AI Ethics Lab, 2020), looking at which principles these documents laid out more often (Jobin et al., 2019; Fjeld et al., 2020; Tidjon and Khomh, 2022; Zeng et al., 2018; Hagendorff, 2020; Khan et al., 2022), or criticising their vagueness (Ryan and Stahl, 2020; Munn, 2022), their inoperability (Tidjon and Khomh, 2022), and their use as a marketing tool (Saltelli et al., 2022)[11].

A few studies examine each principle in detail to see exactly what lies behind these vague terms. For Jobin et al. (2019) and Fjeld et al. (2020), *fairness* is mainly expressed as either the prevention and mitigation of unwanted bias, or non-discrimination. Jobin et al. (2019) also points out that this principle is often linked to the concept of *justice*, which is understood as the respect for diversity, inclusion, equality, and the rule of law. For Fjeld et al. (2020), another important dimension in this principle is also the representativity and high quality of datasets. Rather than conducting a statistical analysis of principles, Evers (2024) focuses on a few charters for which she provides a comparison of the definitions used. She found that for BigTech companies, fairness is an element to be incorporated into machine learning pipelines, with the aim of addressing it through low-cost technological reforms rather than a modification of their business model[12]. She points out that European institutions are also tending to incorporate this techno-solutionist approach in their own documents[13].

Fairness is a broad encompassing term that has often been related to the theory

[11] For more information on these AI ethics critiques, see our background study in Chapter 2, Section 2.2.

[12] This techno-solutionist approach therefore fails to take account of the harms caused by AI systems to the population (Kalluri, 2020).

[13] These results are in line with our analysis in Chapter 2, Section 2.6.

of justice and the rule of law[14]. In recent years, for AI in particular, scholars have seek to understand the interplay between algorithmic fairness and the legal right to non-discrimination. Many legal frameworks have been analysed in that regard, specifically US and EU non-discrimination laws (Weerts et al., 2023; Wachter et al., 2021a,b). Some even seek to understand the lawfulness of machine learning fairness strategies (Kim, 2022; Hellman, 2020).

Recently, another framework has also been of interest to legal scholars working on algorithmic systems: the AI Act. Colmenarejo et al. (2022) look at fairness from an interdisciplinary perspective, where bias, fairness and regulation are discussed along technical, legal and philosophical perspectives. They notably examine the AI Act proposal, noting that it focuses mainly on biases in data sets and relies on the human operator to identify and recognise bias and discrimination. However, they do not provide a thorough analysis of the AI Act's requirements. Deck et al. (2024) notes that the AI Act leaves the judgement of what constitute an illegal discrimination to existing legislation but that non-discrimination law requires a translation from legal requirements to technical fairness requirements.

6.2.2 *Studies on technical standards*

For some years now, the European Commission has been gathering information on AI standardisation state of play. This work is mainly conducted by AI Watch, which is part of the European Commission's Joint Research Centre (JRC)[15], and which carries out a wide range of trustworthy AI-related activities. As part of the forthcoming AI Act, AI Watch has issued various reports (Nativi and De Nigris, 2021; Soler Garrido et al., 2023a,b) to map ongoing standardisation initiatives to the essential requirements for high-risk systems set out in the AI Act. Nativi and De Nigris (2021) calculates an operationalisation level to see how well a standard matches a requirement. For data governance and data quality, the standards identified seem only partially suited to the requirement. Soler Garrido et al. (2023b) provides a more complete description of the relevant standards, albeit with a more restricted list. Additionally, Balahur et al. (2022) summarises the findings of a JRC conference and workshop on data quality.

The academic field has also recently taken a interest in standardisation activities, listing initiatives related to AI (Ziegler, 2020b) or discussing ISO standards (Laux et al., 2024) and harmonised European standards (Golpayegani et al., 2023). However, to the best of our knowledge, these studies have not specifically address fairness standards, nor have they carried out an in-depth analysis of standards. Instead, they frequently limit themselves to enumerating existing initiatives.

6.3 *Methodology*

6.3.1 *Starting point: fairness in ethics charters*

Fairness is a principle that has transcended the world of AI ethics. Although it is not cited in all AI ethics charters[16], it often appears in other forms, such as *justice*, *non-discrimination*, *prevention of bias*, *data quality* or *inclusiveness*[17]. The limits of the principle of fairness are therefore hard to draw.

[14] See notably the work of Rawls (1971).

[15] The Joint Research Centre is an advisory body to the European Commission, which was responsible for monitoring standardisation work in Europe for the AI Act before the AI Office took on this role. For more information, see Chapter 4, Section 4.6.5.

[16] In our database from Chapter 2 (`https://mapaie.telecom-paris.fr/`), *fairness* appears in 263 AI ethics documents out of 436. In the study of Jobin et al. (2019), among the 84 charters studied by Jobin et al. (2019), 68 cite the principle of fairness. It is the second most cited principle after transparency.

[17] Justice is considered to be part of the principle of fairness by Jobin et al. (2019), Fjeld et al. (2020) consider that it includes *prevention of bias*, *data quality* and *inclusiveness*, and *non-discrimination* is considered to be a part of it by both of these studies.

Selection of charters for discussion

For our analysis of fairness in ethics charters, we have chosen to compare the definitions of the four charters most cited in meta-analyses of AI ethics charters (Gornet et al., 2024): the *Ethics guidelines for trustworthy AI* from the High-Level Expert Group on Artificial Intelligence set up by the European Commission (HLEG, 2019a), the Future of Life Institute's *Asilomar AI Principles* (Future of Life Institute, 2017), the *Ethically Aligned Design* framework from the Institute of Electrical and Electronics Engineers (Chatila and Havens, 2019) and the *The Montréal Declaration for responsible AI development* (Université de Montréal, 2018). Added to this list is a fifth charter, the UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2021), which is more recent – and therefore less cited – but which is the largest initiative of its kind with an agreement reached with the largest number of countries.

Among these texts, the Future of Life Institute and IEEE do not have any principle or recommendation directly referring to fairness and thus never define this term. This leaves us with three texts from regional (HLEG, 2019a) and international (UNESCO, 2021) government bodies and the academic world (Université de Montréal, 2018).

Three dimensions of fairness in charters

The HLEG includes in its guidelines a requirement on *"Diversity, non-discrimination and fairness"*. This requirement is broken down into three components: *"Avoidance of unfair bias"*, *"Accessibility and universal design"* and *"Stakeholder Participation"*. In other words, the data sets and algorithms design as well as management practices should allow for the identification and mitigation of discriminatory biases; the AI system should be designed so that anyone could use it regardless of their personal characteristics (age, gender, disabilities, etc.); and stakeholders who will be impacted by the AI system should be consulted before deployment, and their feedback collected afterwards. These three components of fairness according to the HLEG are also found in our remaining charters. The Montréal Declaration includes three principles that directly refer to these aspects (principles 5 to 7): the *"Equity principle"*, the *"Diversity Inclusion principle"* and the *"Democratic participation principle"*. Finally, the UNESCO's recommendations include a principle of *"Fairness and non-discrimination"* (§28-30), which implies that *"AI actors should [...] avoid reinforcing or perpetuating discriminatory or biased applications and outcomes"*, use an *"inclusive approach to ensur[e] that the benefits of AI technologies are available and accessible to all"*, and *"ensure inclusive access [...] and participation in the development of AI"*. In this work, we refer to these three aspects of fairness as (i) "**Bias**", (ii) "**Accessibility**" and (iii) "**Participation**" (the BAP aspects).

Our brief analysis of a few charters shows that fairness is a very broad theme. The different notions of fairness reflected in AI ethics charters are not even always referred to by the term "fairness", but by different concepts related to social justice, democracy and inclusion. This theme therefore goes far beyond the algorithmic fairness generally understood in computer science, and therefore in the technical field of AI, solely as the mitigation of unwanted biases. We will use this general theme and its three dimensions of bias, accessibility and participation as a basis for discussion of how fairness is understood in technical standards.

6.3.2 *Creating a database of fairness standards*

Gathering data

For data analysis, we needed a list of relevant standards. A standard is considered relevant for our list if it meets all the following inclusion criteria:

- **IC1** – It corresponds to our definition of "standard", so a technical document made by experts, regardless of its origin.
- **IC2** – It covers, or intends to cover, AI technologies, either completely or in part (such as robotics and autonomous systems), or a sub-section of AI, in terms of technology or application (such as supervised learning or facial recognition).
- **IC3** – It covers, or intends to cover, one or more of the aspects of fairness defined in Section 6.3.1, namely bias, accessibility and participation.
- **IC4** – It is, or will be, available in English, either free access or via a paywall.

To gather relevant standards, we firstly used the AI Standards Hub database (AI Standards Hub, n.d.), an initiative by the Alan Turing Institute in the UK, which displays a list of more than 400 AI standards, published and under development. These standards are already labelled according to the topics they address. One of these labels is called "bias and discrimination" and yields 23 results. 21 of them corresponded to our inclusion criteria and were added to our list. In addition to this label, we looked at other standards on the platform to see if any of them could match our criteria, which enabled us to add 4 of them to the list. We also checked the online platforms of well-known standards bodies and working groups, such as ISO/IEC JTC 1/ SC 42 (ISO, n.d.f) and CEN-CLC/JTC 21 (CEN-CENELEC, n.d.d), to see if any of their standards were not listed on the AI Standards Hub and discussing fairness. On the ISO and CEN-CENELEC websites, we found respectively 13 and 8 standards under development which will include fairness aspects, with some duplicates from the AI Standards Hub database. Additionally, some ISO standards have also been listed on the CEN-CENELEC website, because they are meant for adoption by European bodies, so they also are duplicates. Finally, 6 standards were added to the list based on our personal knowledge. We thus gathered a list of 42 relevant standards. To make our research more reproducible, we provide a PRISMA-based flowchart of retrieval process in Figure 6.1.

Labelling data

Once we had identified the relevant standards, we retrieved their basic information: the standard name, the name of the organisation that developed the standard, the link to the standard page, whether the standard is open access, a short description of the standard provided by the organisation, whether the standard is published or to be published (TBP), the date of publication and, if relevant, the date of publication of the previous version.

But we also added personal labels depending on the type of standard:

- Horizontal/Vertical (str): whether the standard is horizontal, applies to all AI systems, or vertical, i.e. applies only in certain application domain.
- Domain (str): if the standard is vertical, its domain of application.

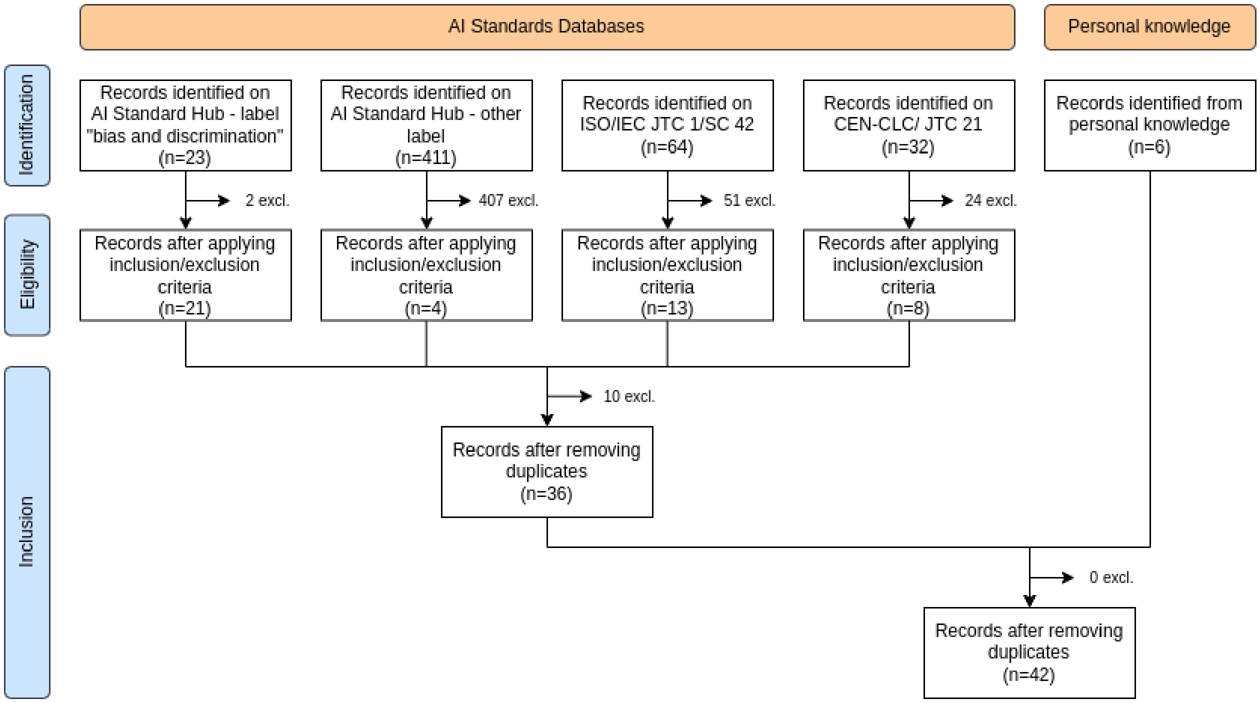

AI Standards Databases
Personal knowledge
Identification
Eligibility
Inclusion
Records identified on AI Standard Hub - label "bias and discrimination" (n=23)
Records identified on AI Standard Hub - other label (n=411)
Records identified on ISO/IEC JTC 1/SC 42 (n=64)
Records identified on CEN-CLC/ JTC 21 (n=32)
Records identified from personal knowledge (n=6)
2 excl.
407 excl.
51 excl.
24 excl.
Records after applying inclusion/exclusion criteria (n=21)
Records after applying inclusion/exclusion criteria (n=4)
Records after applying inclusion/exclusion criteria (n=13)
Records after applying inclusion/exclusion criteria (n=8)
10 excl.
Records after removing duplicates (n=36)
0 excl.
Records after removing duplicates (n=42)

Figure 6.1: PRISMA-based flowchart of retrieval process.

- Technology (str): whether the standard covers directly AI or ML technology, if it covers more largely Autonomous systems and robotics, or specific aspects like data.
- Importance Fairness (int): from 1 to 3, how preponderant is the theme of fairness in the standard. 1: fairness is not directly mentioned but some general ideas are still found in the standard; 2: fairness, or one of its aspects, is one of several dimensions that the standard explores; 3: fairness, or one of its aspects, is the main theme of the standard.
- Information/Requirements (str): is the standard descriptive (information) or prescriptive (requirements). Requirements standards also include a terminology section, but focus on recommending good practice, while information standards simply define contexts, terms or mathematical formulas, but do not give advice on what to do.
- Content (str): if the standard is descriptive, does it provide vocabulary, describe the social context of the technology, give measures or methods for achieving fairness; if the standard is prescriptive, are the guidelines more technical and concern its design, or more non-technical and concern mainly governance and management measures.
- Aspect Fairness (str): what aspect of fairness are covered by the standard, between bias, accessibility and participation.

Standards were annotated with the information available to us. When we did not have access to the standard, we used the description provided by the organisation and, when available, a preview of the standard as provided by the standardisation organisations. When the information we had access to was not sufficient, we added “unknown” to the corresponding label.

A simplified version of this annotated list is presented in Appendix A.3.

To see what the standards on our list contain and what is their advancement in terms of operationalisation, we conduct two complementary analyses. The first one is a quantitative analysis based on our annotated list, where we look at the general distribution of AI fairness standards. The second one is a qualitative reading, where we discuss the content of these standards with relevant examples.

6.4 *Quantitative analysis: the progress of fairness standards*

Figure 6.2 give some basic information about our list of standards.

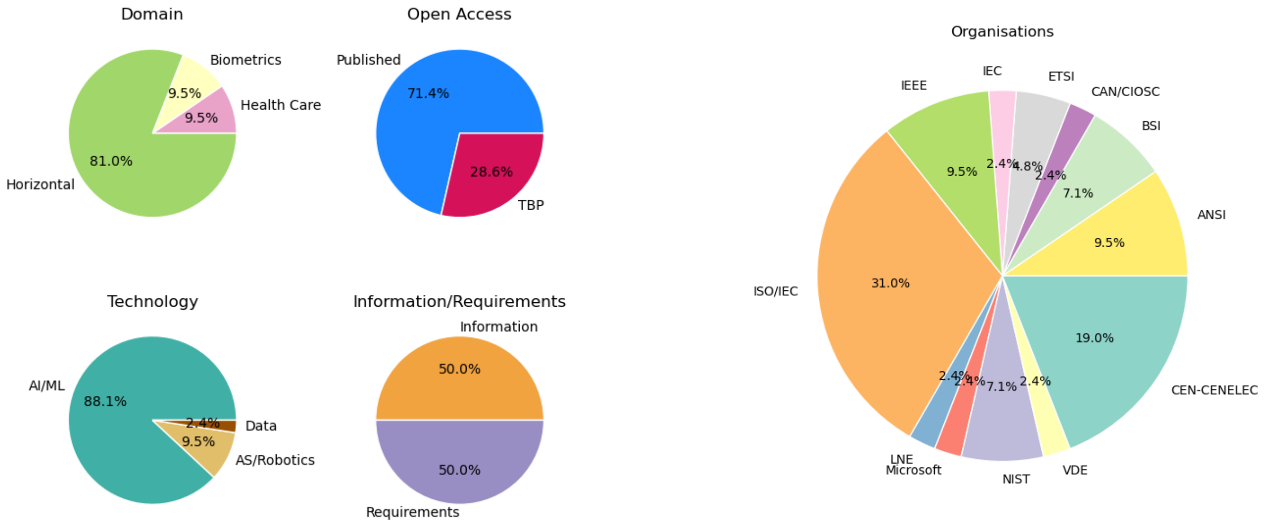


Figure 6.2: Various statistics on our list of standards. On the left, pie charts of domain of application, access, technology and type of standards. On the right, pie chart of the organisations that published the standards.

Most standards on our list are "horizontal" standards – i.e. for all AI systems, and the few "vertical" ones that we have are limited to biometrics and health care. It is worth noting that in the broader list of AI standards that were found on standards databases but did not match our inclusion criteria, other domains are present such as machinery, smart cities, road transports, aeronautics, networks and telecoms, blockchain, etc. But none of these domains seem to address AI fairness. This is understandable as other topics are more relevant to them: for instance, road transports or online networks are more focused on safety and security. About half of the standards in our list are open access, the other half are fee-based. For instance, all ISO and IEC standards need to be purchased, but all IEEE standards are accessible through the GET Program. Most of the standards on our list directly cover AI and machine learning technologies and only one refers only to data. It should be noted, however, that our list does not include data quality standards. Indeed, although data quality can have an impact on fairness, most of these standards refer to data quality only through performance-based objectives and not have fairness considerations in mind. As AI continues to grow, we expect to see standards for specific technologies in the future (computer vision, transformers, foundation models, transfer learning, etc.), but it is understandable that early standardisation efforts will focus on generalisable content. Regarding standards organisations, almost half of the standards on our list come from ISO, the other well-represented organisations being IEEE, BSI, ANSI and NIST. CEN-CENELEC accounts for a smaller percentage, since only two of the standards on our list are developed by them, but it should be noted that many ISO/IEC standards are intended for adoption at European level (Gornet and Maxwell, 2024). Finally, we were surprised to find that almost half of the fairness standards are not merely descriptive, but prescribe a course of action. However, the requirements included in these standards remain very broad in nature, as we will see with the qualitative analysis.

The graph in Figure 6.3 shows us that more and more AI fairness standards are being published. The oldest standard on our list was published in 2016 and deals with the ethical design of robotic systems. We had to wait until 2020 to see the others, but there is now a trend of steadily increasing initiatives that should continue in the future.

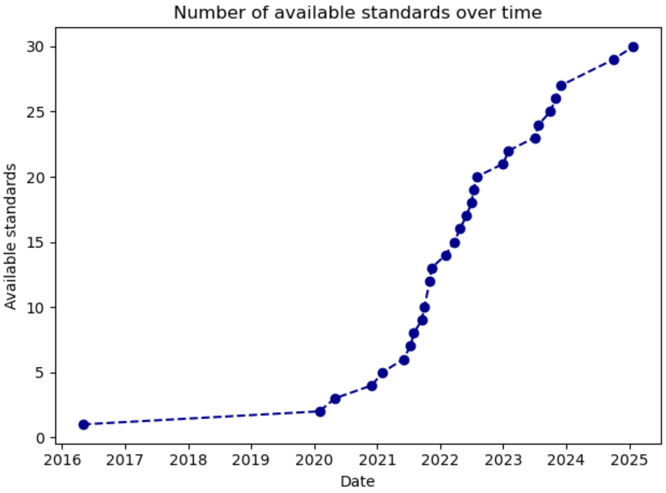


Figure 6.3: Evolution of AI fairness standards over time.

The histogram in Figure 6.4 shows that a large number of standards cover fairness in a shallow manner, either by vaguely incorporating certain concepts without referring to fairness directly (importance of fairness level 1), or by mentioning it among other desirable properties of AI systems in a broader standard (level 2). Few published standards are dedicated to fairness (level 3) even though more of them are in development.

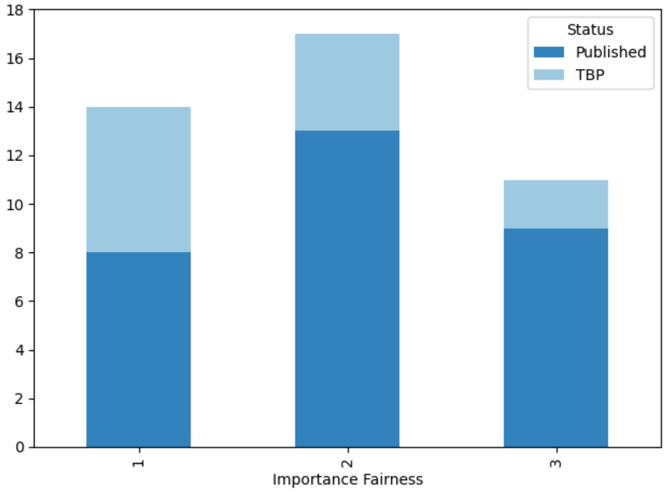


Figure 6.4: Histogram of the number of published standards and standards under development, according to the importance of the fairness theme in these standards.

The histogram in Figure 6.5 shows the BAP aspects broken down for each level of fairness importance (level 1-3). Standards have been counted several times if they deal with more than one BAP aspect. Although we were unable to identify the BAP aspect of all the standards[18], the dominant aspect seems to be, by far, that of bias. Accessibility and participation are addressed, but in fewer standards. This focus on bias is especially true for standards dedicated to fairness (level 3), the only exception being BSI FLEX 236. This trend is not as strong for more general standards, where fairness does not play a central role (level 1). This confirms previous findings in the literature that participatory design and democratic deliberation are almost absent from AI fairness considerations (Weinberg, 2022), and allows us to extend these critiques to fairness standards.

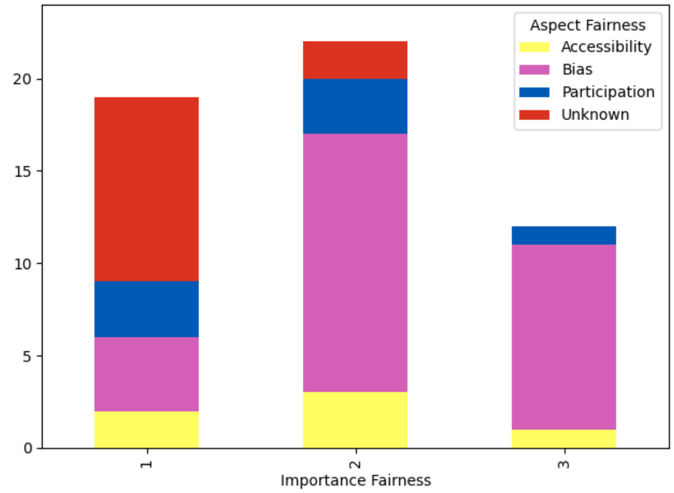


Figure 6.5: Histogram of the number of standards dealing with each aspect of fairness (BAP), according to the importance of the fairness theme in these standards.

In Figure 6.6, we have separated the information and requirements standards into a finer category according to their content. Standards that fell into more than one category were counted several times, once for each category. We find that information standards are dominated by the definition of terms and vocabulary. These are the most basic standards, they are mostly found in more general standards (level 1 and 2), as shown by Figure 6.6. A large proportion of them also define the social context, i.e. the broader ethical and societal issues that might be raised by the development of these technologies. Only one of the standards on our list plans to include methods for the mitigation of biases, but several lay down metrics to identify biases, which is a first step. Standards that define metrics and methods are found exclusively in the category dedicated to fairness (level 3). Regarding requirements standards, more of them include governance requirement than design requirements. Governance requirement are, in our opinion, easier to define, as they can be applied to a wide range of technologies and in different contexts. Design requirements are more technical, thus more technology- and context-specific. As standardisation initiatives continue to develop, we expect standards to progressively incorporate more and more technical requirements in the future.

6.5 *Qualitative reading: standards remain high-level*

We have chosen to divide our qualitative analysis according to the importance of fairness (level 1 to 3), as we have found that this categorisation allows us to compare very similar standards. We only discuss standards to which we have access, unless stated otherwise.

[18] This is mainly due to our limited access to some of these standards.

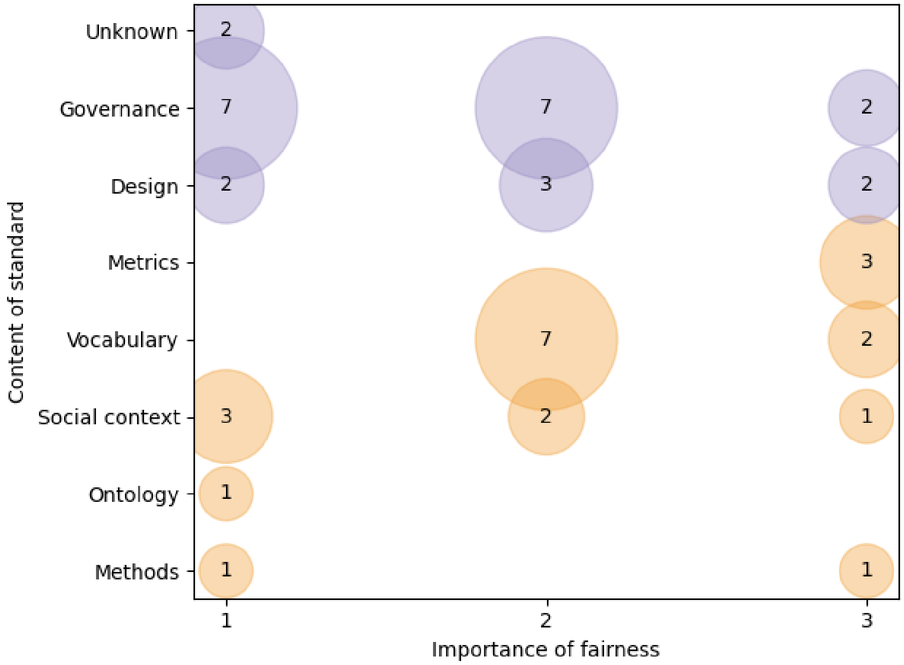


Figure 6.6: Bubble chart of the number of standards in each content category and for each fairness importance category.

Standards on general ethical concerns (level 1)

Level 1 standards deal with ethics in the broadest sense. As a result, in most of them, fairness is not even mentioned, but it is nevertheless a latent element of the various ethical issues. For instance, the CAN/CIOSC 101 standard has an ethics by design part, with one requirement stating that the data should be analysed for unfair biases. However, this fairness element remains minimal.

On a different note, IEEE 7000:2021 is a process-oriented standard that addresses the stages of concept exploration and development. It provides companies with a systematic framework that encourages collaboration with various stakeholders, identifies their needs and ensures the incorporation and prioritisation of relevant ethical values. Fairness and inclusiveness are mentioned in the list of relevant values but are not dealt with specifically. Indeed, the aim of the standard is not to provide specific guidance for these values, but rather to bring stakeholders to define and implement their own design measures and requirements. Similarly, IEEE 7007:2021 provides an ontology of values, including fairness, but does not address the specificities of this theme.

The LNE certification standard for AI processes (LNE, 2021) is a little different, as it is oriented towards auditing and certification. Thus, most of the requirements focus on customer disclosure, including *"levels of diversity, non-discrimination and equity"*. A whole section is devoted to data quality control, in which a study must be carried out to assess biases and their impact, and apply *"debiasing algorithms"* if necessary.

Standards on trustworthiness (level 2)

We call level 2 standards ″trustworthiness″ standards, because fairness is one of many aspects that these standards address. "Trustworthiness" is a term widely used in ethics charters to reference a number of ethical properties that AI systems must meet. This name was chosen because level 2 standards resemble these ethics charters, listing desirable properties of AI systems, including fairness, but not necessarily going further that these charters into technical details.

Depending on the standard, fairness is either briefly mentioned, or is one of the central elements of the standard. Standards where fairness is only briefly mentioned, include only a short section on fairness in which basic definitions of relevant terms are given. Usually, it simply contains a definition of "bias", as in ANSI/CTA 2089, sometimes accompanied by a short (and incomplete) categorisation of biases, as in ETSI SAI 004. ANSI/CTA 2096, although mainly focused on the definition of bias, includes a small recommendation section, stating that datasets should be checked and updated, and that AI developers should test the system for bias, followed by a list of protected features that could be tested.

Some standards also use fairness as an objective in an overall risk management approach. This is the case of ISO/IEC 23894:2023 which includes a fairness objective in its annex A. ISO/IEC 23894:2023 explains how to assess the impacts of AI systems by distinguishing between (1) impact assessment for individuals, including *"potential bias impact"*, *"potential impact on fundamental rights"*, *"potential fairness impact"*, and *"mitigating controls around unwanted bias and unfairness"*; and (2) societal impact assessment, including how a system *"amplifies or reduces pre-existing patterns of harm to different social groups"*. On the other hand, ISO 42001:2023 on AI management system, is divided into several parts: the core of the document resembles any management standard[19] and defines general requirements such as the need to carry out risk assessments, while annexes are more specific to AI issues and detail measures and processes to be put in place to ensure this risk management. Annex A defines objectives, notably on data management, including data acquisition, data sets quality, data provenance and data preparation, while Annex B provides guidelines for implementing these objectives, specifying in particular, for data quality, that *"the organization should consider the impact of bias on [...] system fairness and make adjustments as necessary"*.

Other level 2 standards go further than risk management, where fairness is only a small part, and offer more detailed content on fairness. For instance, ETSI ENI 018 offers a comprehensive categorisation of biases. Additionally, ANSI/CTA 2090 not only defines bias, but also details the ways in which bias can occur, and sets out requirements for developers to avoid bias, such as determining whether the data is already pre-processed, as well as carefully splitting and manipulating data sets.

Some level 2 standards follow a principle-based scheme, inspired by ethical charters, but with a more operational implementation, such as the VDE VCIO AI trustworthiness characterisation (Peylo et al., 2022), the NIST AI Risk Management Framework (NIST, 2023a) and the Microsoft Responsible AI standard (Microsoft, 2022). VDE's and Microsoft's standards both contain detailed design requirements. Common requirements include identifying demographic groups that may be at risk, defining metrics and evaluating the data and the system. However, while Microsoft's standard goes into more technical detail, such as the need to collect data to bridge eventual gaps, document impact factors and publish information, work with domain experts and with members of groups at risk, and use bias mitigation techniques; VDE is more general about the process, requiring to assess trade-offs with fairness, ensure workers' protection and evaluate the system's sustainability.

[19] Such as ISO/IEC 27001:2022 for information security management systems or ISO/IEC 14001:2015 for environmental management systems.

Standards dedicated to fairness (level 3)

Finally, among our standards dedicated to fairness, only one deals primarily with accessibility, and not bias: BSI FLEX 236. Although this standard is not limited to AI and deals with how to develop inclusive standards through user data, it contains many accessibility and inclusive design requirements that are applicable to AI, for example taking into account all potential users and stakeholders. According to the standard, data must be tested for representativeness, taking target users into account, data collection must be carried out carefully to avoid bias, etc. The result is a comprehensive fairness standard, but one that would need to be adapted to better meet the specific issues related to AI.

On another note, NIST SP 1270 and ISO/IEC 24027:2021 both seek to define and categorise bias. NIST SP 1270 distinguishes between *human biases*, *systemic bias* and *statistical bias* and develops social and technical considerations, describing well-known problems in AI fairness, such as taking into account socio-technical factors, interaction with humans or spurious correlations. On the contrary, ISO/IEC 24027:2021 primarily seeks to define terms around bias. The standard gives a list of different types of biases, separated into *human cognitive biases*, *data bias* and *bias introduced by engineering decisions*, as well as metrics to identify these biases, such as equalized odds, equal opportunity, demographic parity and predictive parity. The standard then elaborates on ways to treat unwanted biases, such as including trans-disciplinary experts, measuring the representativeness of samples, using bias mitigation algorithms, conducting internal testing or monitoring the system after deployment. In addition, ISO/IEC 12791:2024, to which we do not have access, appears to build on ISO/IEC 24027:2021 by further developing the part on the treatment of unwanted biases, defining techniques to address unwanted biases in the training, and presenting mitigation measures throughout the life-cycle.

NIST documents also include the Face Recognition Vendor Test Part 8 (Grother, 2022), which goes beyond previous NIST publications on the topics which only summarised the performance of face recognition algorithms on different fairness benchmarks. Indeed, Part 8 defines *equity measures*, either taken from the specialised literature or created from scratch. These metrics differ from current tests carried out in academic literature, which are still often limited to comparing false positives and false negatives. This work has probably influence ISO/IEC 19795-10 (2024a) on quantifying biometric system performance variation across demographic groups, to which we do not have access.

Finally, IEEE 7003-2024 was perhaps one of the most anticipated fairness standard of these past few years, as it was the most cited in standards analyses (Nativi and De Nigris, 2021; Soler Garrido et al., 2023a,b) before its publication in December 2024. Although IEEE 7003-2024 is, once again, a standard focused on bias, it is the most comprehensive standard on fairness to date. The standard includes requirements for “bias consideration”, to enable a provider to establish preliminary thinking about which bias is wanted or unwanted. It also defines documentation requirements in the form of a so-called “bias profile” to record how bias has been considered all throughout the life-cycle. The standards also gives guidance on how to identify potential impacted stakeholders. These are used in further documentation requirements on “data representation”, which include the sources and types of the data, the context of use and the representativeness as regards the different stakeholders. IEEE 7003-2024 also includes a section on bias-related risk and impact assessment. Finally, a last part on "evaluation" aims at assessing bias in the design and output of an AI system.

6.6 Discussion

6.6.1 The paradox of AI fairness standards: technical documents turned into ethics charters

The trend towards broad and open-ended standards began long before AI, notably with software standards (Choi, 2024). Today, this situation is being repeated with AI, as shown by the large number of level 2 standards that we have called "trustworthiness" standards, in reference to their inspiration from principle-based ethics charters.

As a result, standards dealing solely with fairness are quite rare and they provide mainly information, such as definitions of terms and metrics. Requirements are only defined in more general standards on trustworthiness or ethical concerns. "Trustworthiness" standards discuss a whole range of desirable properties of AI systems, such as fairness, but also transparency, human oversight and so on. These standards give guidance and good practices for the development of AI, such as keeping data up to date, analysing data for biases or carrying out risk assessments[20]. However, they rarely go as far as to define implementation methods or testing procedures. They do not give concrete technical methods that an AI developer could follow, nor do they define evaluation criteria to assess the results of an AI system. This could be explained by the fact that most standards today are horizontal, which prevents them from going into the details of implementation.

There is therefore a paradox with AI standards: they are presented as the subsequent evolution to ethics charters, but in reality, they fail to be more technically precise than their ethical counterpart. But there are reasons to this generality of AI standards. On the one hand, despite their technical dimension, criteria resulting from ethical principles such as fairness or transparency are often considered as criteria to be set by the legislator (Baeva et al., 2023). Standardisation organisations often lack the necessary legitimacy to set thresholds for these criteria, as they are closely linked to the protection of fundamental rights, which does not fall within the field of expertise of these institutions (Gornet and Maxwell, 2024). They have therefore focused on creating standards that are more open-ended, not too restrictive and easy to comply with. This makes it possible both to reach agreement in a field where the state of the art is evolving rapidly, and to avoid the responsibility of taking arbitrary choices on highly sensitive questions.

On the other hand, these broad standards, similarly to AI ethics charters, are often ineffective because they lack precision. Indeed, due to the lack of verifiable criteria, they can be followed by companies without them changing their practices, merely documenting risks and verifying basic steps without guarding against any harm. Experts have therefore called for more precise standards, focusing on specific tasks and providing quantitative attributes (Choi, 2024).

The degree of precision required from AI standards will depends on what they are intended to do. Indeed, in the specific case of standards used to comply with a regulation, such as harmonised standards for the AI Act, having technical requirements that are too generic can also create loopholes in the law.

[20] Balayn et al. (2023) speaks of a "checklist" culture, in which standards fit perfectly. For a discussion on how technical considerations permeate the sphere of ethics, see Chapter 9, Section 9.2.1.

6.6.2 *From fairness to data quality and bias mitigation*

The question of what standards should contain is closely related to the question of who should draft them and for what purpose. For instance, fairness standards tend to focus on bias mitigation as it is a topic that is both well-documented in the field of machine learning and scrutinised by the general public and the media. On the contrary, accessibility and participation are topics that are unfamiliar to computer scientists working in AI. The way in which standardisation actors view their own field therefore steers standards in a certain direction. However, this reduced scope of fairness standards risks pushing them into the same framing and formalism traps as the overall field of fair machine learning research (Selbst et al., 2019).

In addition, recent standardisation efforts have focused on supporting the new European legislation for AI: the AI Act. Although fairness is mentioned in some of the AI Act recitals (rec. 27, 74, 94, 110), it is not a property directly required by the legal text. However, some of its related concepts are required, such as data representativeness (Art. 10(3)), as well as the detection and mitigation of possible biases (Art. 10(2)(f)&(g)). To enable the testing of data and biases, providers of high-risk AI systems are required to put in place *"data governance and management practices"* (Art. 10(2)), which will rely on harmonised standards. This focus on data is explained in Recital 67, according to which *"high-quality data"* is vital to ensuring both the performance of an AI system, and to make sure that it does not become a source of discrimination[21]. The other aspects of fairness – Accessibility and Participation, are also found in the AI Act to some extent[22], but no standard is requested on these aspects.

This reductive translation of fairness to data quality and bias mitigation in the AI Act partly explains the limited scope of fairness standards that we witness today. We expect this trend to continue in the years to come as the legislation will gradually enter into force. However, at the same time, the number of fairness standards will also grow. We can therefore expect to see more standards focused on fairness (level 3), as well as more vertical standards that may go deeper into technical details[23].

6.7 *Limitations and future work*

The greatest difficulties in our work was (1) to define the scope of AI fairness, (2) to define what a fairness standard was. Indeed, both "fairness" and "standards" can be interpreted either in a restrictive or broad sense, encompassing different elements. For fairness, we started from three aspects identified in ethics charters, but we recognise that there are many other admissible definitions and many other aspects, and that classifying into different categories implies a certain degree of subjectivity.

For standards, the generally accepted definition can be interpreted in different ways. We chose to emphasise that the organisation of origin was irrelevant, as long as the document was both technical and emanated from experts in their field. But both aspects of that definition are already challenged by some existing AI standards. The technical nature of standards has given way to broader themes, including governance requirements, social and ethical issues. But if any expert can write a standard, and if that standard is not even always technical, any document about AI, describing what AI is or prescribing what to do when developing or using AI,

[21] It should be noted, however, that in reality, data is only one of the many possible sources of biases in an AI system (Suresh and Guttag, 2021), and that even models trained on perfectly balanced datasets can still have biases (Albiero et al., 2020; Gwilliam et al., 2021).

[22] For accessibility, we can cite for instance Articles 16(l), 50(5) and 95(2)(d)&(e), on inclusive design, accessibility for people with a disability, etc. For participation, we can cite Articles 40(3), 56(3), 58(2)(b), 62(1)(d) and 95(2)(d), on the diverse representation of stakeholders in standardisation, sandboxes, codes of practice and codes of conduct, as well as diverse development teams for AI systems.

[23] We discuss in Chapter 9, Section 9.6, why we believe AI standards should be vertical.

could be considered a standard. In our study, most standards come from recognised entities such as ISO or IEEE, the only exception is the Microsoft's Responsible AI standard (Microsoft, 2022) which comes from the industry. Indeed, Microsoft's document differs from common ethical guidelines and charters in that it is more operational. But even among recognised entities, it is hard to decide what constitutes a standard. Even ISO differentiates between international standards identified only by their number, and technical reports, identified by the letters TR in front of their number, such as ISO/IEC TR 24027:2021. However, these documents are all listed in standards databases and are therefore included. On the other hand, the NIST AI Risk Management Framework is listed on the AI Standards Hub, but other NIST documents that we included are not, even though they are detailed enough to be used directly by companies. On the contrary, we excluded technical reports that are not directly operational, such as the AI Watch's Data quality requirements (Balahur et al., 2022). We recognise, however, that this is a difficult line to draw, and that future work could focus on a finer categorisation of documents.

Additionally, there is a large number of data quality standards referenced in AI standards databases, and while we believe such standards can serve a fairness objective, they did not fit our inclusion criteria because they only referred to performance-oriented goals and did not have fairness considerations in mind.

Finally, we believe this work could benefit from a broader study in the future. One way would be to compare standards with external fairness research. Another way would be to extend our quantitative analysis to other aspects of AI ethics, such as mapping standards to other principles defined in ethical charters. We believe our work is a first step in bridging the gap between mostly non-operational ethical principles and implementable standards, but there is still a long way to go to bring these two worlds together.

6.8 Conclusion

We can conclude our analyses with two major statements. First of all, information standards mostly provide definitions and categorisations of bias that are common in the fairness literature. When they define the social context of a technology, the issues raised are also well known, and when they define metrics, they are mainly the best-known, with the exception of biometrics standards, which seem to be rather upstream in the literature. In other words, these standards simply aim to compile the state of the art.

Second of all, requirements standards can be divided into two categories – although some of them deal with both aspects: those primarily concerned with governance, calling for the implementation of risk management practices, improved communication, etc.; and those aimed at the system developer, defining the design guidelines to be respected. Both aspects are complementary and we believe that they are both necessary to achieve fairness. However, we raise two concerns regarding these requirements. First, it is necessary to look beyond the "bias" aspect of fairness. While it is a criticism we can level at all AI fairness standards, it is particularly evident in the requirements standards. The requirements focus primarily on the need for representative datasets and bias mitigation, which we believe is a necessary step, but fails to take into account all the harms the system will continue to cause if it is not tailored to everyone and does not allow for the inclusion of various stakeholders in the development phase. Secondly, with regard to the more technical requirements, while they define good practices and can serve as a sanity check for AI developers, they do not constitute guidelines on the specific methods to be applied in a given context and remain largely general. We believe that, as standards progress, these recommendations will become increasingly detailed. However, they must not become a disempowering tool, not taking the decision for the developer, but enabling her/him to make the right decision in the given context. Notably, we did not see in our list any performance-based standards that would attempt to set a threshold for certain fairness metrics. We believe that such a standard could be extremely dangerous, as it would push AI providers to do the bare minimum and would end up slowing down progress in the field of fairness. We therefore call on standardisation experts to be cautious about the content of the standards they develop, as the role of standardisation in our society continues to grow. Future standards, especially those supporting legislation such as the AI Act, will likely decide which fairness practices are adopted on a large scale.

Chapter 7
Regulation through standardisation: the legitimacy issue of European AI standards

Where does this chapter come from?

The aim of this chapter is to examine the problems with the current standardisation system and how they are reflected in the drafting of AI standards. It is complementary to Chapter 8 in that it draws similar conclusions, but from a study of the literature, whereas Chapter 8 focuses on fieldwork and interviews.

Section 7.3 extends and updates the second section[1] of the following article:

> Gornet, M. and Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, 13(3). `https://doi.org/10.14763/2024.3.1784`

The rest of the Chapter is an original contribution. It was released online as a standalone work:

> Gornet, M. (2024c). Regulation through standardisation: the legitimacy issue of European AI standards. `https://hal.science/hal-04785178`

Chapter's abstract

In December 2023, the European institutions reached a political agreement on the AI Act, a new regulation on artificial intelligence. The AI Act will require providers of high-risk AI systems to test their products against harmonised standards (hENs) before affixing a European Conformity (CE) mark to allow AI products to circulate freely on the European market.

However, the current EU standardisation system faces a number of problems, such as the business model for selling standards, the EU's lack of sovereignty in its own standardisation processes and the possible emergence of conflicts of interest with industry, which raises questions about the legitimacy of the standardisation bodies. Is this system sufficient to protect us from the harms caused by AI?

In this chapter, we examine the existing problems with the standardisation system in Europe and how they are being accentuated in the context of the AI Act. We notably analyse the 5 March 2024 CJEU decision and the respective opinion of the Advocate General in the Public.Resource.Org case and show that there is a growing discontent on the business model of standards. We also examine the international and industrial influences on European standards and show that the tight schedule

[1] The section initially named "The status of harmonised standards in EU law".

imposed by the Commission with the AI Act is likely to intensify them. In addition to their lack of democratic processes, the European Standardisation Organisations (ESOs) do not have the competence to properly address fundamental rights issues, although they are encouraged by the European Commission to diversify their field of expertise. Transparency and contestability of standards will play a crucial role if the European Commission is to avoid the last resort of developing standards itself.

7.1 Introduction

In April 2021, the European Commission revealed its first draft for the future regulation laying down harmonised rules on artificial intelligence (AI)[2], also known as the AI Act (European Commission, 2021e). The text proposed a legal framework to regulate AI systems and laid down requirements that they should meet. The AI Act has since been adopted by the European institutions and will gradually apply across the European Union from 2025 onwards[3].

The AI Act is not the first law on digital technologies in Europe, it follows, notably, the adoption of data protection regulations such as the General Data Protection Regulation (GDPR) in 2016 (European Parliament and Council, 2016b), the Data Governance Act (European Parliament and Council, 2022c) in 2022, and the Data Act (European Parliament and Council, 2023b) in 2023. The Digital Markets Act (DMA) (European Parliament and Council, 2022) and the Digital Services Act (DSA) (European Parliament and Council, 2022a) were also adopted in 2022 for the regulation of online platforms[4]. However, the AI Act takes a different route from these texts, choosing to draw inspiration from European product safety rules. In particular, AI systems will require a conformity assessment that will be based on harmonised standards (hENs[5]), i.e. technical specifications drawn up by European Standardisation Organisations (ESOs) and possessing various legal properties, such as generating a presumption of conformity with the legislation. This conformity assessment procedure will then lead to the European Conformity (CE) marking of the AI product, a seal affixed to show compliance to EU regulations.

However, hENs and CE marking are tools which have already been examined by the academic literature and face numerous criticisms. In particular, the lack of transparency, the involvement of the industry and the legitimacy problem faced by the ESOs weaken the European standardisation system and call into question the suitability of standards as a regulatory tool. In this chapter, we analyse all these issues and look at how they translate into AI standardisation.

We first recap in Section 7.2 the role that standards will play in the AI Act. In Section 7.3, we look in more detail at the status of hENs in EU law, from the James Elliot case (CJEU, 2016a) to the recent 5 March 2024 decision from the Court of Justice of the European Union (CJEU) in the Public.Resource.Org case (CJEU, 2024b). Although hENs have legal effects under European law, they are not freely accessible. In Section 7.4 we show that standards can be seen as a delegation of power from the Commission to the ESOs, but a delegation that lacks control and supervision. We further show in Section 7.5 that the ESOs are subject to international influences, even for standards meant to support European law, which calls into question European sovereignty[6] in standardisation. In Section 7.6, we show that this problem is likely to accelerate with AI because of the dependence of the AI Act on hENs and the tight schedule imposed by the European Commission. The ESOs also face

[2] A definition of AI is given in Chapter 1, Section 1.1.4.

[3] For a detailed timeline of the AI Act entry into force, see Chapter 3, Section 3.5.6.

[4] To know more about the European legal landscape for digital technologies, see Chapter 3, Section 3.3.

[5] For more information on the reasons of the choice of acronym, see box in Chapter 3, Section 4.6.2.

[6] A "sovereign" state or entity, is one that enjoys autonomy and governs itself.

criticisms as regard to the participation of the industry, which we develop in Section 7.7. In addition to the usual criticisms about their lack of transparency, with AI, the ESOs will now have to deal with fundamental rights issues for which they are not equipped. We show in Section 7.8 that these concerns ultimately create a lack of legitimacy that the ESOs struggle to overcome. Finally, in Section 7.9, we show that the European institutions are in fact highly critical of the current standardisation system and are planning for alternatives to the ESOs' standards. In particular, in the context of AI, common specifications will provide a safety net in case standards fail.

7.2 The AI Act strongly relies on standards for conformity assessment

7.2.1 The AI Act is based on product safety

The AI Act is part of the New Legislative Framework (NLF) for product safety[7]. Under the NLF, European legislation[8] does not directly define technical specifications, but rather sets out the "essential requirements" that products must meet, leaving providers and manufacturers some flexibility as to the means of achieving compliance (CEN, 2019). One of the main tool of the NLF is harmonised standards (hENs) (European Commission, n.d.k) which define the technical requirements that would enable a product to comply with the essential requirements set out in a specific product directive or regulation. EU legislation sets what goals to reach, and hENs define how to reach them (Hernalsteen and Kohler, 2022). A harmonised standard is only one possible way to comply with a legal requirement (European Commission, 2022d, p.50) and is thus intended to be voluntary like any other standard, as stated by Article 2(1) of Regulation 1025/2012 (European Parliament and Council, 2012). However, it is in practice the most important pathway for compliance.

hENs are developed by one of the three ESOs: the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardisation (CENELEC), or the European Telecommunications Standards Institute (ETSI). If a directive or regulation needs to be supported by hENs, the European Commission issues a standardisation request to one or more ESOs, describing the main topics the standards should cover[9]. Once the standards have been drafted by the ESOs and approved, they are generally published in the Official Journal of the European Union (OJEU)[10], therefore benefiting from some legal effects such as a presumption of conformity. This means that when a manufacturer of a product covered by NLF legislation has followed hENs, the product will automatically be considered compliant with the corresponding legislation.

hENs are particularly useful in conformity assessment procedures, where a manufacturer or third party assesses compliance with legislation and, if hENs are used to demonstrate compliance, procedures can be simplified. Once the conformity assessment has been completed, the manufacturer affixes a CE mark to the product, indicating that it complies with current legislation and can be sold in the EU[11] without restrictions. The AI Act uses this product-based approach, as AI systems considered "high-risk" must go through this conformity assessment procedure and receive a CE mark before being put on the European market.

[7] For more information on the NLF, harmonised standards and CE marking in Europe, see Chapter 4, Section 4.6.

[8] Directives and Regulations.

[9] Not all standards developed by ESOs, are hENs, only those following a request from the Commission (Art. 2(1)(b)&(c), Reg. 1025/2012).

[10] Not all harmonised standards are cited in the OJEU. Some might be requested by the European Commission to address standardisation gaps, without supporting a specific legislation (Hernalsteen and Kohler, 2022).

[11] More precisely in the European Economic Area (EEA).

7.2.2 *Private organisations draft harmonised standards*

European and international standardisation organisations are private associations that are tasked to develop technical standards. They are composed of experts which have signed a service contract with a National Standardisation Body (NSB), another private entity gathering experts from the same country. Experts can come from private companies, research institutes, public establishments, or work on their own behalf. Anyone can apply to join a NSB to take part in standards development and committee voting, generally in exchange for a membership fee[12], paid by the expert's institution. However, because of the cost of sending experts to work on standardisation, industry is often over-represented (Morikawa and Morrison, 2004)[13].

Once experts are part of their NSB, they can ask to join the working groups at European or international level. This includes the three ESOs and the three international standardisation bodies: the International Organisation for Standardisation (ISO), the International Electrotechnical Commission (IEC) and the International Telecommunication Union (ITU)[14].

The various NSBs and standardisation organisations at international or European levels may have different legal statuses[15] and business plans. However, they generally depend on income from membership fees, the sale of standards and, in some cases, a certification assessment service. NSBs and standards bodies also work together to ensure that standards developed at international or European level are adopted and sold at national level. With a few exceptions, access to these standards is therefore subject to a fee, and standardisation organisations retain copyright over the standards, even though they are written by external experts. Standards can even have different prices depending on which NSB sales them. For Gestel and Micklitz (2013), this is proof that they are trying to make the most of their position in the market.

7.2.3 *Conformity assessments are mostly be carried out by providers themselves*

Products that fall under the NLF, such as high-risk AI systems, need to undergo a conformity assessment procedure. To this end, manufacturers can choose to rely on any technical specifications, including hENs. For certain products, the conformity assessment must be carried out by a notified body, which are often private entities. But for many products, recourse to a notified body is not even necessary. For AI, a third-party audit will only be required for high-risk systems covered by existing NLF regulations and for biometric systems under certain conditions. But for the "new" areas of application introduced by the AI Act in Annex III, where the use of AI means it is automatically considered high-risk, such as education, employment or justice, providers will simply have to rely on internal control[16].

The entire compliance control chain, from the development of standards to support legislation to the auditing of systems against these standards, is therefore carried out entirely in the private sector. The European institutions have only the power to approve and supervise the work of these private entities. This large weight of the private sector may prove to be a flaw in the AI Act. Indeed, in the sectors listed in Annex III of the AI Act, companies will send experts in standardisation organisations to develop standards and will themselves carry out the conformity assessment using these same standards. This creates a loophole insofar

[12] Some NSBs do not require experts to pay a membership fee, for instance in Austria.

[13] See Section 7.7.

[14] For more information on the functioning of these different standardisation organisations, see Chapter 4, Section 4.5.

[15] Some of them are non-profit organisations, other are for-profit organisations.

[16] For more information about the different conformity assessment procedures under the AI Act, see Chapter 4, Section 4.7.2.

as providers of high-risk AI systems, who are supposed to be regulated, will instead hold the keys to regulatory compliance with the rules they have established for themselves (Wachter, 2024).

7.3 *Harmonised standards are not freely accessible*

7.3.1 *Harmonised standards have legal effects and can be considered part of EU law*

The legal significance of technical standards in the EU has grown, because regulations cannot be understood without their relevant standards, making them *de facto* binding (Cantero Gamito, 2018; Everson et al., 1999). Some consider that the development of technical standards has entered a stage of "juridification" (Schapel, 2013), a term taken up by the recent Opinion of the Advocate General in the Public.Resource.Org case (CJEU, 2023, §29). hENs are now regarded as a form of implementing acts (Tovo, 2018).

A number of cases have involved the analysis of the scope of hENs. The Fra.Bo SpA v Deutsche Vereinigung case (CJEU, 2012a) showed that hENs can have de facto mandatory effects, due to the presumption of conformity granted to them that renders any other means of achieving compliance more costly and time-consuming. Additionally, the Court of Justice of the European Union (CJEU) held in the James Elliott Construction Limited v Irish Asphalt Limited case (CJEU, 2016a) that hENs form part of EU law due to these legal effects. The last case to date, Public.Resource.Org Inc. And Right to Know CLG v European Commission – also known as the Malamud case, named after the founder of Public.Resource.Org – examined whether hENs could be subject to copyright protection. After an initial ruling by the General Court (CJEU, 2021), the relevance of the claim to copyright protection was re-examined in an appeal. To this end, the Advocate General, in his 22 June 2023 Opinion, conducted a detailed analysis of hENs (CJEU, 2023). The Court delivered its judgement on the appeal on 5 March 2024 (CJEU, 2024b).

Even if Regulation 1025/2012 considers hENs to be voluntary in theory, as there are other ways to demonstrate compliance, in practice it is difficult if not impossible for manufacturers to choose a different avenue. Recourse to hENs is thus quasi obligatory for economic players if they want to stay competitive (Van Elk and Van der Horst, 2009). Another advantage is that the presumption of conformity reverses the burden of proof, since the company does not have to prove that it complies with the legislation, as this is automatically presumed. If a manufacturer chooses not to comply with hENs, the onus is on him to prove that his product complies with the legislation, which represents a huge commercial risk that no manufacturer would take (CJEU, 2023, §42). As noted by the Advocate General in the Public.Resource.Org case appeal, the whole architecture of the EU standardisation system presupposes that all actors use hENs (§47). According to the Advocate General, there are no realistic alternatives, because ESOs are too focused on hENs development to propose other standards and there is no financial incentive for other private actors to compete with them (§48).

7.3.2 *The commercial operating mode of ESOs is at odds with the legal scope of harmonised standards*

The Public.Resource.Org decision (CJEU, 2021) involved two non-profit organisations who requested access to several hENs, referenced in the OJEU but whose full text was not public and behind a paywall. In 2021, the Commission refused to grant them this access on the basis of the first indent of Article 4(2) of Regulation 1049/2001. This article lists the exceptions to the free access of the EU institutions documents, and states that access can be refused *"where disclosure would undermine the protection of commercial interests [...] including intellectual properties [...], unless there is an overriding public interest in disclosure"*. A first judgement was made on 24 July 2021 by the General Court, in favour of the Commission. In their appeal, the organisations asserted that the General Court erred in incorrectly assessing the copyright protection of hENs, since hENs are part of the law and cannot be copyrighted, and if they were allowed copyright protection, free access to the law would take precedence over copyright protection. While the European Commission claimed that the European standardisation system cannot function without paid access to standards, the two non-profit organisations considered that this does not prevail over the right of access to these standards. According to the Vademecum of the European Commission (2015), hENs are only a means to support the implementation of legislation. In the Public.Resource.Org case appeal, the Advocate General questioned this claim, affirming that they are more than a simple aid and are actually an *"essential tool"* for the correct implementation of EU legislation (§33-36). One of the Advocate General's conclusions is therefore that, due to the heavy reliance of EU legislation on hENs, the effectiveness of the legislation is compromised in the absence of a publicly accessible version of these standards. hENs are indeed considered by the Advocate General to be "indispensable" for enforcing the corresponding EU legislation, thus, the public cannot exercise their rights if they do not have access to hENs (§46-47). To ensure that everyone can have the possibility to know the law and respect it, every act, including hENs, should respect the principle of transparency and right of access to documents, recognised by the Consolidated Version of the Treaty on European Union (European Commission, 2012b, Art. 1§2, 10(3), 11(2)&(3)) as well as the Charter of Fundamental Rights of the European Union (European Commission, 2012a, Art. 42). This is at odds with the operating mode of ESOs that usually charge for access to technical standards and keep the intellectual property of all their standards.

In addition, the Grand Chamber found that *"[harmonised standards] may be necessary for [individuals] to verify whether a given product or service actually complies with the requirements of [a] legislation"* (CJEU, 2024b, §82), emphasising the principles of transparency and openness to which democratic institutions are subject under EU law (§83). In this regard, the Grand Chamber agreed with the non-profit organisations, concluding that there was indeed an overriding public interest in the disclosure of these standards. The initial judgement by the General Court was set aside and the European Commission will need to give access to the four requested harmonised standards. This judgement, however, does not seem to question the copyright protection of hENs, as stated by CEN-CENELEC (2024a). Yet, it is unclear if this decision entails an automatic publication of hENs in the OJEU or a simple disclosure upon request (Soroiu, 2024).

Update March 2025

The judgement resulting from the Public.Resource.Org case seemed to have been interpreted differently by the various parties involved. In August 2024, CEN-CENELEC set up a web portal providing access to the four harmonised standards requested in the case, but not to other harmonised standards[17]. This portal was removed a few days later.

On 27 September 2024, the European Commission published a decision amending Regulation 1049/2001 of 30 May 2001 regarding public access to European Parliament, Council and Commission documents, to include access to standards that have been cited in the OJEU, whether international standards, such as ISO or IEC standards, or European standards, such as CEN-CENELEC standards. Right after, the Commission introduced its own solution for providing access to harmonised standards: access was granted on individual request and was read-only on an online platform. The interested individual could request access to the harmonised standard via the Electronic Access to Commission Documents (EASE) platform[18]. The request was analysed by the Commission and, if granted, the person would receive an e-mail with a link to the read-only online platform. Access to harmonised standards is also currently possible through a new web portal set up by CEN and CENELEC which provides links to national versions of harmonised standards[19].

On 6 December 2024, ISO and IEC brought an action against the Commission before the EU Court of Justice, to *"protect international standards from unauthorized disclosure"*[20]. ISO and IEC are asking the Court to annul the Commission's decision and to order the Commission to compensate them for the costs. They plea that the Commission has made *"manifest errors of assessment and lack of reasoning"* in the interpretation of the Public.Resource.Org case (OJEU, 2025). According to the two organisations, the scope of the judgement was limited to EU harmonised standards and *"there is no overriding public interest [...] in favour of the disclosure of IEC and ISO International Standards"*. Furthermore, they claim that the Commission has breached Article 17 of the Charter of Fundamental Rights of the EU which safeguards the right to property. They consider that the disclosure of international standards will *"seriously undermine the commercial interests and the valuable intellectual property of IEC and ISO"*. Finally, the two organisations consider that the Commission has infringed Regulation 1049/2001 regarding public access to documents, and specifically Article 4(1) which obliges institutions to refuse access to a document where disclosure would undermine the protection of international relations, citing in particular the WTO agreements.

At the time of writing this thesis, the judgment is pending and it is unknown whether access to harmonised standards will be maintained in the future.

[17] A visualisation of this old web portal is possible through internet archives: `https://web.archive.org/web/20240820014810/https://has.standards.eu/`.

[18] `https://ec.europa.eu/transparency/documents-request/home`

[19] `https://harmonized.standards.eu/`

[20] According to a joint statement published on social media by journalist Luca Bertuzzi: `https://www.linkedin.com/posts/luca-bertuzzi-186729130_update-iec-and-iso-issued-a-joint-statement-activity-7275445905367928832-4Vd3/`.

7.4 There is a lack of control over standards

7.4.1 The Commission is responsible for political choices while the ESOs are responsible for technical choices

Today, hENs are published in the OJEU under the letter L, for legislation, where previously they were published under C, for information and notice (§9). As confirmed by the various CJEU decisions, hENs are the equivalent of a legally binding regulation, even though they are developed by institutions – the ESOs – without any democratic accountability. In reality, hENs are developed under the direction of the Commission, the executive branch of the EU that could be seen as the politically responsible author of the standards.

The James Elliott case (CJEU, 2016a) found that the Commission has significant control over the procedure of drafting and considered hENs as constituting acts of the institutions of the EU. To begin with, the Commission has the role of requesting hEN through standardisation requests. These requests take the form of Commission Implementing Decisions. However, despite the formality of the legal Decision, some authors argue that because of the private nature of the ESOs, the request is more akin to a contractual relationship (Cuccuru, 2020). Yet, not only does the Commission request hENs, it also supervises the drafting and adopts them. After the draft harmonised standard has been proposed by the ESOs and before publication in the OJEU, the Commission is empowered to send back the document to the ESOs for modification if the draft does not comply with the request. Ultimately, publication in the OJEU depends on acceptance by the Commission. The cycle of an hEN thus starts and ends with the Commission. This led the Advocate General in his Opinion on the Public.Resource.Org case appeal (CJEU, 2023) to conclude that the Commission has the power to transform a preparatory document into an act that forms part of EU law (§28). The Advocate General further advises that the Commission should be seen as the institution adopting hENs and that ESOs are only preparatory bodies (§17).

The European Commission itself has declared (European Commission, 2022e) that more power needs to be transferred from the ESOs to the Commission. One way of achieving this would be to allow the Commission to draw up technical solutions directly, as an alternative to the hENs drawn up by the ESOs. The AI Act acknowledges this possibility with "common specifications", an alternative to standards drafted by the Commission itself[21].

7.4.2 Standardisation as a delegation of power

The term "delegation" of power is never used directly in CJEU ruling, as the court prefers to say that the Commission *"entrust"* the development of harmonised standards to private bodies (CJEU, 2016a). But in various cases, Advocate Generals are more direct, speaking of a *"'controlled' legislative delegation in favour of a private standardisation body"* (CJEU, 2016b) or a *de facto* transfer of competence to private associations (CJEU, 2012b). It is therefore clear that, in the context of harmonised standards, the European Commission delegates part of its power to the ESOs to draw up these legal acts. However, the CJEU sets narrow limits for the delegation of powers under the Meroni doctrine (ECJ, 1958b,a): only clearly defined powers, which are the result of an express delegation, and which are subject to strict reviews,

[21] See Section 7.9.4.

can be delegated. But this doctrine might no longer hold since the European Securities and Markets Authority (ESMA) case (Scholten and van Rijsbergen, 2014), where the Court loosened these requirements by concluding that delegation of power is possible when there is an adequate judicial supervision (CJEU, 2014).

Despite the Commission involvement, democratic oversight of hENs is still lacking, as neither the European Parliament nor the Member States have a right to veto standards. Additionally, the Commission's right to refuse publication of a hENs is burdened by technical limitations and human resources costs that prevent it from carrying out a comprehensive examination (Ebers, 2022). As a result, the Commission can refuse to publish standards in the OJEU, but this assessment is only based on a strict comparison between the contents of the standard, the SR items and the requirements of the law (Ebers, 2022). Scholars therefore believe that the delegation of power from the Commission to the ESOs in the case of standardisation is permitted under EU constitutional law, if the absence of control prior to the publication of standards can be compensated by an ex post judicial review (Eliantonio, 2017).

7.4.3 *Contestability of standards might not be enough*

Although the Commission is politically responsible for standards, its control prior to the publication of harmonised standards is limited. But even after publication, according to Ebers (2022) it is unlikely that the Court would be willing to rule on the validity of a harmonised standard, either in an annulment action[22] or a preliminary ruling procedure[23]. Even if it were, the CJEU is unlikely to review and invalidate its substantive content: its jurisdiction would be limited to reviewing whether the Commission made an error in making the decision to publish a harmonised standard in the OJEU (Ebers, 2022). This pushes Ebers (2022) to affirm that hENs are *"currently in essence immune from judicial review"*.

However, this is not entirely true. The current Regulation on standardisation (European Parliament and Council, 2012) provides Member States with the power to oppose hEN when they do not provide a sufficient level of conformity with essential requirements (Art. 11 Reg. 1025/2012). This opposition is called "formal objections". France used this mechanism, for example, for standard EN 50566:2013 in July 2014. The standard was supposed to cover Article 3 of Directive 1999/5/CE on hertz equipments (European Parliament and Council, 1999) – which was later repealed by the Radio Equipment Directive (RED) directive (European Parliament and Council, 2014). At the time, France estimated that the recommended standards for the specific absorption rate (SAR), a measure of the rate of energy absorbed by the human body, did not correspond to the actual conditions under which phones and other mobile devices were used. Indeed, while the standard recommended measuring SAR at a maximum distance of 25mm, it was more common to have a device within 5mm of the body. Measures conducted by the French National Frequencies Agency (ANFR) confirmed these claims (ANFR, n.d.) and the Commission subsequently decided, by an implementing decision to change the content of the standard. The reference of standard EN 50566:2013 is now published in the OJEU with a warning explaining that for trunk SAR measurements, *"a separation distance of not more than a few millimetres may be used"* (European Commission, 2016a). The changes was justifies as regard to the safety objectives of the corresponding Directive. However, such objections are unlikely to succeed in the event of a challenge to fundamental rights. To the best of our knowledge, this has, at least, never been the case.

[22] Per Article 264 of the Treaty on the Functioning of the European Union (TFEU) (European Commission, 2012c).

[23] Per Article 267 TFEU.

7.5 European standards face international influences

7.5.1 The necessary alignment between European and international standards

The processes followed by standardisation organisations are inherited from international agreements, that are adopted at a larger scale than Europe. Indeed, in all their activities, ESOs and international SDOs follow the "core principles for standardisation" (WTO, 2000), and the "code of good practice for the preparation, adoption and application of standards" (WTO, 1995, Annex 3), set out by the World Trade Organisation (WTO). Both the principles and the code of practice are part of a broader agreement, called the "WTO Technical Barriers to Trade" (TBT) agreement (WTO, 1995). The WTO code of good practices is notably said to be applied by over 200 standards-setting bodies world-wide (WTO, n.d.).

The aim of the TBT agreement is to *"ensure that regulations, standards, testing and certification procedures do not create unnecessary obstacles"* (WTO, n.d.). As such, the WTO TBT agreement pushes the use of international standards as a basis for regulation, even if it leaves a degree of flexibility with respect to the choice of standard and the manner of its use (Wijkström and McDaniels, 2013). The agreement notably encourages countries to recognise each other's procedures to avoid the duplication of standards and tests that have to be carried out from one country to the other. ISO and IEC play a big role as they notably compile which country participate in the agreements, and their terms and definitions are directly used in the TBT agreement[24]. Under this agreement, regional standards and regulations are therefore forced to align with international standards, particularly those of ISO and IEC.

7.5.2 International standards have priority over European ones

This approach of convergence of standards encouraged by the TBT agreement pushes standardisation organisations to put in place mechanisms to favour international standards. In Europe for instance, the ESOs can collaborate with international SDOs and directly adopt their standards to make them "European standards", even though these international bodies are not made up entirely of European members. For instance, the Vienna and Frankfurt agreements, concluded respectively between CEN and ISO, and between CENELEC and IEC, facilitate the exchange of information between the bodies and avoid duplication of work (ISO and CEN, 2016b; CENELEC, 2017). This collaboration extends to the adoption of standards, since ISO and IEC standards can be incorporated into the catalogue of European standards by ratification by CEN-CENELEC[25]. At present, nearly 35% of CEN publications come from ISO, and 81% of CENELEC publications come from IEC (CEN-CENELEC, 2024). These figures seem to have increased in recent years compared with previous studies (Cuccuru, 2019). Furthermore, this adoption extends to hENs: 28% of all CEN deliverables which are offered for citation in the OJEU[26] come from ISO, respectively 69% for CENELEC and IEC (CEN-CENELEC, 2024). The Vienna agreements state that ISO standards should be adopted by CEN "if possible" without change (ISO and CEN, 2016a). This could compromise the process of re-examining the suitability and legitimacy of a standard for a given task, and lead ESOs to adopt ill-suited international standards[27].

These agreements also allow expertise to be shared between SDOs, with members participating in several organisations and sometimes collaborating on the develop-

[24] Annex 1 of the TBT on "terms and their definitions" is partly extracted from the ISO/IEC Guide 2 (1991).

[25] See the interactions between the different SDOs in Chapter 4, Figure 4.1

[26] i.e. harmonised standards which benefit from a presumption of conformity.

[27] This push to adopt international standards as they stand is criticised by the experts we interviewed in Chapter 8.

ment of standards which could be adopted by both ISO and CEN, respectively IEC and CENELEC. In the Vienna agreements, this possibility of developing "common ISO/CEN standards" requires that the work that result from the collaboration be subject to *"parallel approval procedures in both organisations"* (ISO and CEN, 2016a). This parallel development notably requires one organisation to take the lead in standardisation work. However, the agreements stipulate that an ISO-lead standard is the preferred option, and that a CEN-lead standard is only possible if it is supported by committee members who are from ISO. As far as hENs are concerned, ISO and IEC standards take precedence where they exist, unless it can be proved that the Commission's request cannot be met by standards issued by these international (Cuccuru, 2019). In order for CEN to take the lead in the development of hENs, ISO has to *"agree that the project should advance under CEN lead based on clear documentation that the ISO committee cannot meet the EC Standardisation Request"* (ISO and CEN, 2016a).

This collaboration makes the composition of international standards organisations even more relevant to European issues, since their standards are likely to become hENs. Indeed, international organisations encompass a much wider range of stakeholders than just the European players. For instance, while a large proportion of ISO members come from Western Europe, almost half come from elsewhere in the world, particularly Asia and North America (Morikawa and Morrison, 2004)[28]. With regard to AI in particular, the secretariat of ISO/IEC SC 42 – the sub-committee in charge of AI standards[29] – is based in the US, as it is held by the American National Standards Institute (ANSI) (ISO, n.d.f).

Furthermore, it is not only ISO/IEC standards that can be technically adopted by ESOs and become hENs. To meet regulatory expectations in accordance with the request for standardisation, ESOs may rely on existing products from any organisation, provided that there is *"no suitable ISO, IEC or ITU deliverable or activity likely to meet the need"*. This means that standards developed entirely outside Europe, such as the US standards developed by the American National Standards Institute (ANSI) or the US National Institute of Standards and Technology (NIST), could eventually become hENs (CEN et al., 2021). However, this is not the preferred option, and contrary to ISO/IEC standards, American standards such as the AI Risk Management Framework (AI RMF) (NIST, 2023a), will likely not become hENs.

7.5.3 *Some European stakeholders do not have a say in the development of international standards*

All SDOs are structured around NSBs, as experts must first register with their national body in order to take part in standardisation discussions in international and European working groups. However, at European level, there are some exceptions to this system as some specific stakeholders have the right to participate in the ESOs working groups without registering first with a NSB. This is notably the case of associations representing societal interests – i.e. every non-business organisations – as well as associations representing the interests of small-medium sized enterprises (SMEs). CEN-CENELEC calls such stakeholders "Partner organisations" (CEN-CENELEC, 2021). Some of these organisations are also entitled to funding from the European Commission. They are called Annex III organisations (A3Os) as the criteria for receiving funding are listed in Annex III of Regulation 1025/2012 (European Parliament and Council, 2012). These include organisations

[28] It should be noted however, that this study looks at participation in ISO in the late 90s. Participation has probably evolved since then.

[29] To understand the structure of ISO/IEC and CEN-CENELEC as regard to working groups on AI, see Chapter 5, Figure 5.1.

representing European SMEs, consumers, environmental interests and social interests (Annex III, Reg. 1025/2012). Four organisations were identified by the Commission to represent these interests: (i) Small Business Standards (SBS) representing SMEs; (ii) European Association for the Co-ordination of Consumer Representation in Standardisation (ANEC) representing consumers; (iii) European Environmental Citizens' Organisation for Standardisation (ECOS) representing environmental interests; and (iv) European Trade Union Confederation (ETUC) representing workers (Ernst & Young, 2020).

The added value of the participation of these stakeholders in European standardisation is recognised by both the European Commission and the ESOs. In particular, CEN-CENELEC has stated that they help *"provid[e] expertise"*, *"contribut[e] to balanced representation"* and *"legitimis[e] the standardisation system"* (CEN-CENELEC, 2017). However, these A3Os, and more generally the partner organisations, do not necessarily have the right to participate in the work of ISO and IEC, since these SDOs remain based on the participation of national bodies. This means that these organisations have no say in the development of international standards, even if they are subsequently adopted as European standards (Cuccuru, 2019). The adoption of international standards as European standards as a result of the TBT and Vienna agreements helps to converge the work of different bodies, but nevertheless restricts the voice of civil society, which is more present at EU level (Cuccuru, 2019).

7.5.4 *Even when standards are developed within the ESOs, European actors are not the only ones to participate*

Questions of international influences are also raised when it comes to which countries participate in the discussions, even when the standards are developed exclusively by the ESOs. Indeed, the ESOs welcome various NSBs beyond the borders of the European Union. Different countries can benefit from different levels of inclusion in ESO standardisation work, from simple observer to full member with voting rights[30]. In particular, in addition to the twenty seven members of the EU, seven other countries are full members of CEN-CENELEC. This includes, for instance, the British Standards Institute (BSI), the UK's NSB. As there is no difference between these seven countries and the rest of the NSBs, experts can also occupy leadership positions. Examples include leading a standard project in working groups[31] or hosting the secretariat of technical bodies within the ESO. With the BSI, the UK is the third country with the most secretariats at CEN-CENELEC, behind Germany and France (CEN-CENELEC, 2024).

In addition to the country leading the discussions, the question of the nationality of the companies taking part is also of the utmost importance. While experts participate as individuals when they are registered in a NSB, they may represent the interests of transnational companies, sometimes even based outside Europe. This is particularly true of American BigTech companies, such as the GAFAMI[32], which have branches in European countries and whose experts are therefore directly involved in European standardisation work in the ESOs.

[30] For more information on CEN-CENELEC membership, see Chapter 4, Section 4.5.2.

[31] For example, obtaining the role of convenor or editor. For more information on the different roles in standards setting, see Chapter 8.

[32] Google, Amazon, Facebook (Meta), Apple, Microsoft and IBM.

7.6 European standards are in a race for sovereignty

7.6.1 Timing issues are due to a lengthy European standardisation process

Although the European standardisation system is regularly praised in European official communications[33], European institutions have admitted on several occasions that the current system has some flaws, in particular regarding the amount of time it takes to publish standards (European Commission, 2011c). The first criticisms of the slowness of the European standardisation process date back to the EU Commission's "Green paper on the development of European standardisation" in 1990 (European Commission, 1990), and have been repeated ever since (Hesser and Czaya, 2010). The Commission is aware that this long timeline often puts at risk the implementation of laws that are directly supported by these standards and that must subsequently wait for their development. To counter this, the Commission explains having negotiated in the early 2000s with the ESOs to reduce the time taken by internal processes to accept and publish hENs. As such, according to a study conducted by the Commission, the time taken by hENs to be developed went from eight years in 2003 to three in 2008 on average (European Commission, 2011b).

But years after years, the length of development of standards, although getting better, is still cited as an important issue, even as of today[34] The Commission therefore seems well aware that *"the success of the implementation of the European standards will depend on the legitimacy and the efficiency of their adoption processes"* (European Commission, 2011c).

As a result, the adoption of international standards is sometimes presented as the solution to timing issues. The European Parliament notably recognises that *"the broader uptake of existing standards rather than the drafting of new ones would [...] have positive effects on timing"*. It further recommends taking the time to prepare new standards as *"the unduly rushed preparation or deployment of standards creates challenges for all stakeholders"* (European Parliament, 2023c). This injunction to the ESOs is, however, at odds with recent developments in standardisation, where the European Commission is pressing for standards, particularly in the field of AI, to be developed more quickly[35].

7.6.2 There is a gold rush to determine who will draft AI standards

The time constraint of AI standards is their greatest obstacle. To begin with, there is a gap between the speed of deployment of AI-based products and services and the development of standards. This creates pressure on standards bodies to publish standards as quickly as possible, otherwise they run the risk of incorporating technical elements that are already obsolete.

For the moment, European AI standards are lagging behind their international counterparts. Calls have therefore been made for closer links with ISO (NEN, 2021) and convergence with international standards (JBCE, 2021). Indeed, ISO is ahead of the game in developing technical standards for AI: at the time of writing[36], SC 42 already published 31 standards and 36 more are currently under development (ISO, n.d.f). On the contrary, CEN-CENELEC has yet to publish its own standards on AI[37].

There is therefore considerable tensions as regard to which of the European or international bodies will develop the standards that will shape AI in Europe. This

[33] In a communication from 2011, the Commission says that *"European standardisation was extremely successful and one of the driving factors of the creation of the internal market for goods"* (European Commission, 2011c).

[34] In the recent EU strategy on standardisation (European Commission, 2022e), the Commission stated that *"European standardisation must respond to an increasingly rapid innovation pace and needs to deliver standards fast [...]. In particular in new and emerging technologies, the European standardisation system often fails to deliver in a timely manner and hence loses the important 'first mover' advantage through standardisation"*.

[35] For more information, see Chapter 9, Section 9.5.4.

[36] November 2024.

[37] For more information on the current state of EU and international AI standardisation initiatives, with a particular focus on ethics and fundamental rights, see Chapter 5, Section 5.3.

tension is even more important when it comes to standards relating to the ethical aspects of AI systems, where visions may diverge between the European Union and other regions of the world. Some societal stakeholders question whether Europe should trust international standards to incorporate these elements, arguing that there is no guarantee that these standards are consistent with EU rights and values (ETUC, 2021). ANEC has notably expressed its concerns about the adoption in Europe of standards in which non-European countries or companies have participated (Russell, 2021) and calling for the ESOs to address EU values and *"not just adopt international standards which might not reflect our values and principles"* (Giovannini, 2021). They propose that the Commission should clarify whether a harmonised standard can be entrusted to ISO, or whether it should be developed within the ESOs (Russell, 2021).

This ongoing competition is reinforced by the tight schedule imposed by the AI Act. European standards need to be quickly available for the AI Act to be properly enforced. Indeed, the text relies heavily on the subsequent publication of these technical documents, which currently do not exist (Pouget, 2023). Although the Commission has requested standards by April 2025, many believe that this deadline is impossible to meet (Perarnaud, 2023). Yet, if the ESOs fall too far behind the deadlines imposed by the European Commission, they may be tempted to adopt international standards that are already ready, instead of developing their own. The urgent need for European AI standards therefore raises issues of both correct implementation of the regulation and European sovereignty.

7.6.3 Europe needs its own definitions

Ideally, it is better to align the definitions of different frameworks that deal with the same topic. For instance, EU institutions and the OECD have converged on definition of AI, so has the Council of Europe, and the US are in the process of doing the same. Aligning definitions enables the various stakeholders to communicate and implement rules effectively. For legislation and standards, alignment is all the more important as different definitions could lead to poor implementation or regulatory loopholes. However, in the context of the Vienna agreements, alignment of European standards with definitions emanating from international stakeholders and decided at ISO level raises questions of sovereignty, because if a European standard is contradictory, it should be withdrawn.

Yet, different terms may be understood differently in international standards and in European regulation such as the AI Act. For instance, the notions of quality management, risk, representativeness, transparency, intended use, accuracy, etc., have a specific meaning in the AI Act. But they are also defined in standards that may not be aligned with the definition given in the regulation. For instance, in the AI Act, the risk are considered to the safety, health and fundamental rights of individuals. On the contrary, in international standards, risks are understood as business risk for companies. In addition, quality management in ISO is a well established notion since ISO 9001 and refers, like risk, mostly to business expectations. Yet, under the NLF in European law, it is considered as one of the elements to check for conformity assessments and has its own characteristics defined in the corresponding European regulation, such as in Article 9 of the AI Act.

Similarly, for Europe, if the AI Act does not give a definition of trustworthiness, the HLEG guidelines do: a trustworthy AI system is a system that is ethical, lawful and robust both from a technical and social perspective. If this definition could be shifted in further works, notably by European standards, it nevertheless reflects Europe's desire to be open to more dimensions than just a technical one. On the contrary, trustworthiness is defined in international standards as *"the ability to meet stakeholders expectations in a verifiable way"* (ISO/IEC, 2020).

But this lack of alignment is even more worrying when it comes to technical terms such as data representativeness, transparency or accuracy. Indeed, European standards, if they want to become harmonised and be published in the OJEU need to respect the vocabulary and meaning of the law, which is not always aligned with the definitions of international standards. European standards find themselves in a place where they cannot align at the same time with international standards and European regulation and where both the international experts also present at JTC 21 and the European Commission are pushing for alignment.

7.7 *The strong participation from the industry can create conflict of interests*

7.7.1 *The industry is over-represented in standardisation*

Standardisation organisations have a large history of being dominated mainly by the private sector (Brunsson and Jacobsson, 2002), particularly large companies and certification organisations (Baeva et al., 2023). In ISO for instance, the largest stakeholder group is the industry, and the private sector as a whole accounts for around two-thirds of all participants (Morikawa and Morrison, 2004)[38]. Indeed, private actors have always had more incentive to participate in standardisation, which explains why this activity first developed outside government spheres (Hesser and Czaya, 2010). Standardisation activities enable companies to increase their market share and influence, avoid costs by being informed of future compliance requirements at an early stage, and save time and money on product testing (de Vries et al., 2009). This dominance of the private sector is also explained by the resources required to participate (Baeva et al., 2023), whether in terms of the cost of full-time employees working on these subjects, the fees sometimes required by the NSBs or the cost of travelling all over the world to attend meetings.

Due to their capacity to deploy large resources, private organisations have emerged as a driving force in standardisation, and have ultimately proven to be better at producing effective standards than public entities (Yates and Murphy, 2019). This composition also gives standardisation organisations access to beneficial industrial expertise (McFadden et al., 2021), an essential competence for the development of technical requirements related to product safety. McFadden et al. (2021) even believe that more industrial expertise, such as that of international SDOs, would be beneficial to ESOs, whose stakeholder group is more diverse.

However, not everyone is happy with the important role of the industry in standardisation, with some feeling that key stakeholders are not sufficiently represented (Werle and Iversen, 2006). With regard to AI standardisation in particular, Edwards (2022) deplores the fact that those affected by AI have no role to play in standardisation or certification processes. This has prompted some scholars to call for greater participation of civil society in standardisation, to counterbalance the weight of the industry and bring more legitimacy to standardisation organisations (Baeva et al., 2023).

[38] The study was conducted in TC 207 on Environmental management. "Industry" represented 32%, "Consulting" 18%, and "Standards Organization", which are also private bodies, represented 21%. Note however, that the study dates from the late 90s.

7.7.2 *Standards can lead to economic war between companies*

This industry-led composition can lead to competition between companies to impose their standards, which is sometimes to the detriment of standardisation and the legitimacy of the standards bodies. The most well-known example is perhaps the competition, in the early 2000s, between the Office Open Extensible Markup Language (OOXML) standard and the Open Document Format for Office Applications (ODF) standard.

In 2008, the American BigTech company Microsoft proposed to translate an ECMA (the former European Computer Manufacturer's Association) standard – the ECMA-376-1 OOXML standard – into a new ISO standard: ISO/IEC 29500:2008. However, a few years earlier, in 2006, international organisations ISO and IEC already adopted a standard for document format: ISO/IEC 26300:2006, also called the ODF standard, supported among others by IBM, another American tech giant. Several ISO members, except Microsoft and ECMA, expressed serious reservations as regard the adoption of a new standard that would contradict the ODF one (Blind, 2011). After initially failing to pass, the draft standard was proposed in a fast-tracking standardisation process and managed to obtain the 75% approval criterion of the votes cast by participating members of JTC 1 on Information Technology, and was subsequently published in addition to the existing ODF standard (ISO, 2008).

The standardisation processes in the OOXML case have been widely criticised, with strong accusations of procedural flaws and irregularities, including vote manipulation and vote buying, committee stuffing and intervention of public authorities[39]. The debate became so heated that IBM reportedly threatened to leave the standards body (Kirk, 2008).

The case of the ODF and OOXML standards is therefore often presented as an example of economic competition between opposing companies, IBM and Microsoft (Blind, 2011). But this further shows that international standardisation organisations, such as ISO, allow dominant companies, such as Microsoft, to exert undue influence on the standardisation processes (Kirk, 2008). This influence could threatens the legitimacy of standardisation organisations.

7.7.3 *Standards capture political and social aspects*

Standards are strongly associated with power (Busch, 2011). Since their early days, European standardisation has been very political, as differences on the political level were reproduced on the standard one (Hesser and Czaya, 2010). As pointed out by Solow-Niederman (2024), *"standards have politics"*, they are neither objective nor neutral. They are the result of political steering by both public and private powers. Since standards have a strong regulatory power, specifically in Europe, they attract many different stakeholders who all have an interest in influencing AI rule-making.

However, in technical fields, the industry has all the power, due to the economic dominance of BigTech companies (Almada and Petit, 2023). In AI in particular, as people become increasingly dependent on these technologies, the corporate resources that BigTech can deploy give them considerable power (Whittaker, 2021). This economic power is reflected in standardisation, where the industry is over-represented, which, according to Werle and Iversen (2006) give them the power to steer the choices of standard organisations towards their preferences. This industry-

[39] A list of irregularities are compiled in a Wiki held by the association <NO>OOXML: `http://noooxml.wikidot.com/irregularities`.

led composition therefore raises risks of regulatory capture and conflicts of interest, since industrial stakeholders are drafting the very same laws by which they will be governed (Bryson, 2022). Regulatory capture is defined by Dal Bó (2006) as *"the process through which special interests affect state intervention"*. According to Mitnick (2011), the industry "captures" decision making *"so that what regulators decide [...] is what industry prefers they decide"*. This risk of regulatory capture is accentuated by the strong lobbying activities of these private actors (Büthe and Mattli, 2011).

The risk of capture is particularly high in standardisation, where decisions are taken by "consensus", i.e. everyone has to agree for a decision to be adopted[40], which confers considerable power on all the stakeholders and risks steering technological development towards those who are most represented (Yates and Murphy, 2019). This work of operating has led (Schmidt and Werle, 1998) to assert that, although standards are called "technical", they are often constructed in processes that are not technical.

But beyond their processes, the question of what they should address is also strongly debated. Some believe that standardisation is wrongly moving away from scientific and technical issues to embrace social issues that require political consensus (European Council of Engineers Chambers, 2021). Similarly, according to EDRi (2022), standards should not try to take decisions that require democratic scrutiny or legal interpretation. Some go so far as to accuse the European standards organisations of playing politics with standards (McFadden et al., 2021). Standardisation is therefore often seen as a way to bypass traditional law making (Abbott and Snidal, 2009). The standardisation of AI is no exception as, according to Ebers et al. (2021) it is *"not a matter of purely technical decisions. Rather, a series of legal and ethical decisions must be made, [...] which require a political debate involving society as a whole"*.

7.8 The ESOs face a legitimacy problem

7.8.1 The ESOs lack transparency and democratic accountability

From the inside, standardisation organisations generally think that their processes resemble deliberative democracy, as they work on the basis of "consensus". From the outside, however, they look more like a technocracy, since expert knowledge is usually a condition for being included in the discussion (Yates and Murphy, 2019). For instance, the ESOs, as numerous SDOs, have agreed to the follow the WTO principles (WTO, 2000), which are meant to guide international standard-setting organisations, including the principles of "transparency" and "openness"[41]. Even if these principles can simply be considered as good governance, the WTO has been called out for not respecting its own principles, given that its meetings are closed to the public (Charnovitz, 2005). The same applies to standardisation organisations, as working group meetings, and subsequently ongoing work on AI standards, are reserved for expert members who have registered with a national body, or who are part of a partner organisation. Transparency is therefore understood in the standardisation world as transparency within organisations, not transparency of the organisation itself.

In addition to their lack of transparency, the ESOs also lack democratic accountability. Even if responsibility for issuing the hENs is shouldered in large part by the Commission, ESOs that develop the standards are governed by private law, lacking the democratic legitimacy of the Commission and the other EU institutions. As

[40] For a discussion over what consensus means and how decisions are taken, see Chapter 8, Section 8.4.3.

[41] The other principles are "Impartiality and Consensus", "Effectiveness and Relevance", "Coherence" and "Development Dimension".

private law bodies, the transparency and accountability of the ESOs are necessarily different from those of public authorities (Eliantonio and Cauffman, 2020). The delegation of power that operates under the Meroni doctrine, from the Commission to the ESOs, is therefore "excessive" according to some (Ebers et al., 2021). With AI, the lack of legitimacy of the ESOs is even more worrying, because of the consequences that the deployment of AI systems could have on society (Wachter, 2024).

7.8.2 *ESOs do not have the competency to deal with fundamental rights*

The legitimacy of ESOs is further challenged by the AI Act, as standards will encompass fundamental rights issues and ESOs lack the expertise to assess them (Veale and Borgesius, 2021). Indeed, ethics and fundamental rights are at the heart of the AI Act discourse, and standardisation initiatives are multiplying in an attempt to address these normative questions[42]. However, it is extremely difficult to standardise ethics and fundamental rights, and some academics believe that it could be almost impossible to translate concepts such as "risks to fundamental rights" into technical criteria without the help of lawyers (Smuha and Yeung, 2024).

Fundamental rights might therefore be outside of the scope of expertise of the ESOs (EDRi, 2022). The engineering experts who sit in standardisation organisations are unlikely to be familiar with human rights jurisprudence, what constitutes interference with fundamental rights and what may be considered necessary in a democratic society (Smuha and Yeung, 2024). There is therefore a need, either to recruit new experts from a wider range of backgrounds, or to train staff working in these SDOs in legal issues (Almada and Petit, 2023). In a previous version of the standardisation request, the European Commission stated that CEN-CENELEC should ensure to *"gather relevant expertise in the area of fundamental rights"* (European Commission, 2022f, Art. 2(1)). This is necessary to ensure the relevance of technical standards with judicial norms, yet it might not be sufficient to guarantee the legitimacy of the ESOs in the establishment of EU legal acts dealing with the protection of fundamental rights.

This lack of legitimacy can be extended to the notified bodies who are in charge of the conformity assessment procedure in certain cases. To have the right to conduct conformity assessments, notify bodies must be accredited in accordance with the ISO/IEC 17011 (2017) standard, demonstrating notably their impartiality and the competence of their staff. While this accreditation justifies their technical knowledge of a specific field, it does not account for their expertise in fundamental rights issues. For the AI Act specifically, many systems will not be audited by a third party and the conformity assessment will be carried out internally. This calls into question the legitimacy of a provider of an AI system to assess the risk of their product to fundamental rights, particularly when this assessment is carried out without external oversight.

7.8.3 *The different types of legitimacy undermined in the standardisation processes*

Legitimacy is often simply defined as to be *"allowed by law"* or *"acceptable"* (Cambridge Dictionary, n.d.). But to Senden (2020), legitimacy is also about the validity and the justification of authority. Scholars often separate legitimacy into three types: "input", "output" and "throughput" legitimacy (Schmidt, 2013). Input le-

[42] For an overview of standardisation initiatives related to ethics and fundamental rights, see Chapter 5, Section 5.3. For an overview of standardisation initiatives on fairness specifically, see Chapter 6.

gitimacy is judged from the effective participation of citizens into policy making and the responsiveness of institutions to their concerns; output legitimacy from the effectiveness of the policy outcomes; and throughput legitimacy from the efficacy, accountability, transparency, inclusiveness and openness to interest consultation of the governance processes. Some warn that these types of legitimacy could be in tension with each other, for example the inclusion of more diverse interests in policy making, and therefore greater input legitimacy, could reduce the efficiency of this policy and therefore its output legitimacy (Eliantonio and Cauffman, 2020). However, a good balance between input, output and throughput legitimacy is necessary for an authority to be considered legitimate.

These three types of legitimacy, input, output and throughput, could all be compromised in standardisation processes (Senden, 2020), and in particular for AI standards. First, standardisation bodies lack some diversity in the interests that are represented, with the industry being over-represented; they therefore lack input legitimacy. This is reinforced by their lack of competence in fundamental rights matter, a crucial aspect of AI standards. This explains why the EU has always strive to include civil society actors in regulation processes: to make them more legitimate (Dunkerley and Fudge, 2004). However, historically, standardisation has been more concerned with its output legitimacy, i.e. having standards that work well (Werle and Iversen, 2006). Output legitimacy is a hard dimension to evaluate as it strongly depends on the content of the standards and the sector in which they are deployed. For instance, in the telecoms sector, where interoperability is necessary for systems to function, output legitimacy is very high (Cantero Gamito, 2018). On the contrary, food safety standards are not the most effective (Bevilacqua, 2020). To assess the output legitimacy of AI standards, we will therefore have to wait until they are published. Nevertheless, scholars worry that the shift towards open-ended quality management frameworks might undermine the effectiveness of AI standards (Choi, 2024)[43]. Finally, standardisation bodies might lack throughput legitimacy because of their design and development processes which lack transparency (Volpato and Eliantonio, 2020).

The overall picture shows a standardisation system that is striving to improve its input and throughput legitimacy by seeking to include more interests in the standardisation process and by adhering to principles that advocate transparency. But despite these efforts, the current standardisation system remains biased and opaque.

7.9 The Commission is forced to find solutions

7.9.1 Including more SMEs and societal stakeholders: and EU priority

CEN-CENELEC states that *"it is a principle of standards activity that all interests affected by the work are taken into account"* (CEN-CENELEC, 2001). Similarly, the EU has always strive to include civil society actors in regulation processes (Dunkerley and Fudge, 2004). However, in practice, large companies still dominates discussions on standardisation, despite smaller companies also having interests in participating (de Vries et al., 2009). The effective participation of European SMEs and societal stakeholders is therefore recognised by European institutions as a major challenge (European Commission, 2011b; European Parliament, 2023c). In particular, the European Parliament (2023c) has called on the European Commission to

[43] We discuss this possibility in Chapter 9, Section 9.3.1.

find better ways to accommodate for SME participation in standardisation.

SMEs are most often excluded because they do not have the necessary financial resources (Cauffman and Gérardy, 2020). Indeed, the main barriers to access to the ESOs are the time required by standardisation work, travel costs and entry fees (European Commission, 2011b). Similarly, some societal organisations such as ANEC or BEUC seek to get more involved (Smuha and Yeung, 2024). However, apart from bigger NGOs, most organisations do not have experience in standardisation (Smuha and Yeung, 2024).

The Commission is therefore making great efforts to diversify participation in standardisation. With the standardisation regulation, it has pushed for the inclusion of societal stakeholders in the ESOs by negotiating a special place for them as associates of CEN and CENELEC, granting them voting rights and a seat at the table (European Commission, 2011b). The Commission also provides funding, in the form of "action grants" and "operating grants"[44], to enable these organisations to send experts. This funding is provided trough the means of different programmes, directed to SMEs, consumer organisations, NGOs and trade unions (European Commission, 2011b).

In the context of AI, this already existing desire to strengthen the participation of the various stakeholders is reaffirmed. Recital 121 of the AI Act notably states that *"a balanced representation of interests involving all relevant stakeholders in the development of standards, in particular SMEs, consumer organisations and environmental and social stakeholders [...] should therefore be encouraged"*, and Article 40(3) further calls on standards setting bodies to *"enhance multi-stakeholder governance ensuring a balanced representation of interests and the effective participation of all relevant stakeholders"*.

7.9.2 *The dual discourse of European institutions regarding international standards*

In general, the ESOs strongly encourage the development of relationships between themselves and international organisations. They recognise, for instance, that *"any interested party from any part of the world can participate in the technical work of CEN and/or CENELEC through different channels"* (CEN et al., 2021). According to Hesser and Czaya (2010), the ESOs are therefore "regional" standards bodies that aspire to become "global players". For some scholars, the expanding boundaries of the ESOs to the rest of the world is a good development and can be seen as a gain in autonomy from the power the European Commission exercises over them (Hudson et al., 2013).

On the other hand, the European Commission and the other European institutions have more mixed opinions when it comes to international standards. Depending on which documents we consider, European institutions convey a dual discourse, sometimes pushing for the adoption of international standards, and sometimes calling European bodies to develop their own standards for the sake of European sovereignty. In 2011, the Commission notably stated that *"European standards should [...], wherever possible, be based upon the internationally accepted standards of ISO, IEC and ITU"*. In the same document, it called for greater cooperation on standardisation with a number of countries, including the US, China, Russia, Japan, India and Brazil (European Commission, 2011c). More recently, the European Parliament has also recognised that adopting international standards *"would allow for easier understanding and implementation"* (European Parliament, 2023c).

[44] See the conditions to receive these special grants at (European Commission, n.d.j).

However, recent communications from the Commission show that the discourse on the role of international stakeholders in European standardisation may be shifting. In particular, in the 2022 EU strategy on standardisation, the Commission recognises that *"in sensitive areas [...] other world regions are taking the lead in international technical committees promoting their technological solutions, which are often incompatible with the EU's values, policies and regulatory framework"*. The strategy therefore sets a clear objective of *"technological sovereignty"* (European Commission, 2022e). This determination to remain sovereign in the field of standardisation is best described by the sentence: *"European standards [...] must be decided by European players"* (European Commission, 2022g). To ensure this goal, the 2022 standardisation strategy notably proposes to tighten the rules for the drafting of harmonised standards. As a result, Regulation 1025/2012 has been amended (European Parliament and Council, 2022b) so that, in order to be eligible for Commission standardisation requests, delegates of the EU NSBs must be those with decision-making power at each stage of the development of the standard. For the Commission, this is a way of avoiding *"any undue influence of actors from outside the EU and EEA*[45] *in the decision-making processes during the development of standards"* (European Commission, 2022g).

7.9.3 *A lack of trust in standardisation: the case of ETSI*

In the 2022 standardisation strategy, the European Commission states its concerns regarding the decision-making processes of the ESOs which gives an *"uneven voting power to certain corporate interests"* (European Commission, 2022e). It therefore calls on the ESOs to *"modernise their governance structures"*. The Commission, although referring to the three ESOs, calls out ETSI in particular as *"multinationals have acquired more votes than the bodies that represent the entire stakeholder community"*.

ETSI is usually less studied in the literature (Volpato and Eliantonio, 2024), but recent works have raised ongoing concerns regarding its governance (Volpato and Eliantonio, 2024; Cantero Gamito and Marsden, 2024; Kanevskaia, 2024; Stanojevic, 2024; Wiegmann, 2024). This interest in ETSI's governance was triggered in particular by the fact that the European Commission left it out of the standardisation request for the AI Act (European Commission, 2023a), which was only addressed to CEN and CENELEC. This exclusion was seen as a sign of the Commission's desire to limit foreign influence on AI standards (Bertuzzi, 2022b; Perarnaud, 2023).

ETSI was created specifically to give Europe a voice in ICT standardisation and to ensure that these standards could be support EU legislation (Cantero Gamito, 2018). However, ETSI differ from CEN and CENELEC in its membership structure. While CEN and CENELEC are based on NSBs, ETSI allow for direct representation of industry and national governments. It also has a "one table approach" where representatives from different sectors sit at the same table, while CEN and CENELEC are separated into different sectoral groups. This approach is seen as on of the reasons for the success of ETSI, whose standards are broadly accepted by market players (Cantero Gamito, 2018). Another difference lies in the business model of the three ESOs. While ETSI includes funding from international stakeholders, CEN and CENELEC still rely on EU funding (Bertuzzi, 2022b).

Although ETSI has undertaken a number of reforms in response to criticisms (Volpato and Eliantonio, 2024), some concerns remain about its legitimacy and procedural safeguards. If this mistrust were to last, it could be detrimental to the European

[45] European Free Trade Association.

standardisation system as a whole (Kanevskaia, 2024).

The case of ETSI therefore shows that the European institutions are cautious regarding the ESOs and their governance model. Like ETSI, CEN-CENELEC could be under scrutiny if it does not properly address its ongoing problems regarding international and industrial influences, the transparency of its processes or its lack of competence in fundamental rights.

7.9.4 *More power to the Commission: the role of common specifications*

The AI Act, although strongly based on compliance with harmonised standards, notably for high-risk AI systems, also provides for an emergency mechanism in case standards fail: common specifications. Common specifications do not date back to the AI Act, but they remain a fairly recent mechanism. They are, for instance, present in other recent European legislation, such as in Regulation 2017/745 on medical devices, also known as the Medical Device Regulation (MDR) (European Parliament and Council, 2017b), and Regulation 2017/746 on in vitro diagnostic medical devices, or more recently in Regulation 2019/881, also known as the Cybersecurity Act, or in Regulation 2023/1230 on machinery products[46]. In all these regulations, to the best of our knowledge, there have only been three cases of common specifications, which apply for certain medical devices (European Commission, 2020a, 2022b,c).

Common specifications are "technical specifications" within the meaning of Regulation 1025/2012, Article 2(4), i.e. technical standards that are not adopted by "recognised" entities such as the ESOs, but are adopted by the Commission by means of implementing acts (Art. 41(1) AI Act). For AI, the Commission can adopt common specifications in four situations: (i) when the standardisation request has not been accepted by the ESOs; (ii) when hENs are not delivered within the deadline; (iii) when the proposed hENs "insufficiently address fundamental rights concerns"; and (iv) when the hENs do not comply with the request. According to the Commission, the purpose of common specifications is to *"ensure that the public interest is served where harmonised standards are absent and insufficient"* (European Commission, 2022e). Common specifications provide for the same presumption of conformity as harmonised standards (Art. 41(3) AI Act), however, they should be repeals once harmonised standards are published on the same topics (Art. 41(4) AI Act). In the case of the MDR, common specifications are essentially mandatory, unless the manufacturer *"can duly justify that they have adopted solutions that ensure a level of safety and performance that is at least equivalent"* (Art. 9(3)).

It seems clear by Article 41 that the Commission could decide to adopt common specifications for AI if standards are late. However, it is still debated in which circumstances the Commission could decide that the proposed harmonised standards are not fit for purpose. Gonzalez Torres and Ali-Vehmas (2024) suggests that a harmonised standard consultant (HAS consultant), who is responsible for assessing the standards to be harmonised on behalf of the Commission before they are published in the OJEU, could for instance refuse a hEN when they consider that it does not respect EU values.

The Commission strongly insist on the fact that common specifications should be *"an exceptional fall back solution"*[47] (rec. 121 AI Act) and should not become the norm. Similarly, for the Parliament, *"this mechanism should only be used in exceptional cases"* (European Parliament, 2023c). The gradual integration of common specifica-

[46] Other examples include Directive 2016/2102 on the accessibility of the websites of public sector bodies, Regulation 2019/1009 on fertilising products, Regulation 2023/1542 on batteries, Regulation 2024/1789 on the hydrogen and gas market and Regulation 2024/1781 on ecodesign requirements for sustainable products.

[47] The 2022 strategy on standardisation (European Commission, 2022e) also uses the term *"fallback solution"* to refer to common specifications.

tions into European legislation could signal the risk of capture from the ESOs (Cantero Gamito and Marsden, 2024), and could be seen as a way for the Commission to put pressure on the ESOs to deliver standards on time and without too much industrial and international influence. As a result, common specifications are not really to the taste of standardisation organisations, which would loose their monopoly on the development of technical standards. The German standardisation bodies have even recommended their elimination during negotiations on the AI Act (DIN and DKE, 2021).

Common specifications are not a miracle solution. First, they would only be drafted *after* hENs have failed. Indeed, for the Commission to have the right to invoke Article 41, the standards must be late, so the common specification process would have to wait until the end of the official deadline. This means that common specifications would probably be developed in a hurry. This is why the Commission recognises that this solution is not recommended in the event of a delay due to the *"technical complexity"* of the standard (rec. 121 AI Act).

Furthermore, common specifications may not solve the problem of the influence of private entities, since the Commission is encouraged to *"cooperate with international partners and international standardisation bodies"* in drafting common specifications (rec. 121 AI Act). For medical devices, the Commission was assisted in the drafting of common specifications by the Medical Device Coordination Group (MDCG), a group of experts from Member States which was established by the MDR (avanti europe, 2024). The MDCG is composed of different subgroups, including a subgroup on standards (European Commission DG SANTE, n.d.).

For the AI Act, it is not clear who exactly would draft these common specifications. While the AI Office is expected to support the Commission in developing these frameworks (European Commission, 2024c, Art. 3(2)(d)), it is not clear whether it should do so alone or with the help of external stakeholder groups. According to Baeva et al. (2023), the Commission could therefore choose to adopt a framework from anywhere in the world. But it could also choose to take the same stakeholders as in the ESOs, simply bypassing traditional standardisation processes. However, it is most likely that, if common specifications were to be drafted, the AI Office would request help from the AI Board, which, similarly to the MDCG for the MDR, was established by the AI Act[48], and also has a subgroup on standards (Bertuzzi, 2024).

The question therefore remains as to whether common specifications could have more legitimacy than standards. Mazzini and Scalzo (2023), two former EU Commission officials, explain that, since they are adopted by implementing acts, common specifications will have to respect the examination procedure whereby member states have to provide a positive opinion for it to pass. On the contrary, the EU Parliament is worried that *"recourse to implementing acts affects the co-legislators' powers of scrutiny"* (European Parliament, 2023c). In its amendments to the AI Act, the Parliament proposed that the Commission should provide justifications before resorting to common specifications (European Parliament, 2023a, amendment 445). However, this element has not been retained in the final version of the text. Furthermore, for Cantero Gamito and Marsden (2024), common specifications would be *"equally flawed as standardisation in terms of input legitimacy"*, for two reasons. Firstly, it is likely that the discussions over common specifications will remain *"largely technocratic and behind closed doors"*. Secondly, the Commission does not have more legiti-

[48] For more information on the AI Board, see Chapter 3, Section 3.5.4.

macy in terms of fundamental rights as it is an executive actor, not a parliamentary one.

Following the 2022 standardisation strategy, the Commission has decided to re-evaluate Regulation 1025/2012 to *"assess whether it is still fit for purpose"* (European Commission, 2022e). A public consultation has been launched and a Commission decision is expected in the fourth quarter of 2024 (European Commission, n.d.h). According to EU officials, the revision of Regulation 1025/2012 could be an opportunity to include common specifications in the list of possible technical specifications to facilitate their use in future European legislation (Mazzini and Scalzo, 2023).

7.10 Conclusion

This chapter has examined the European standardisation system as a whole and the various criticisms it faces. With this work, we provide insights into how these issues might translate into AI standardisation and we show that the problems already raised by the academic literature are likely to accelerate with AI. We notably examine recent case law that has determined the role of hENs in European law, as well as the 5 March 2024 CJEU decision and the respective Advocate General's Opinion in the Public.Resource.Org case appeal. This case law deals a blow to the current standardisation system by recognising that its business model of selling standards might not be adapted to the legal power hold by hENs. Indeed, hENs are to be regarded as EU legal acts and, while the Commission is to be held responsible for the political dimension of hENs, the ESOs are responsible for the technical content. Despite this prominent role of hENs in European law, they are not untouchable. The consultant sent by the Commission to review hENs before they are published could decide that they do not match European goals, and even when published, they can be challenged by Member States. However, these mechanisms are insufficient to challenge standards on fundamental rights questions that they will nonetheless have to address with AI.

In addition to its business model, the standardisation system is also heavily criticised for its governance process. Standards are influenced by international stakeholders even when they are intended to support European law, and industry is still over-represented despite the Commission's efforts to diversify the profiles of participants. With AI, the world of standardisation has embarked on a gold rush to develop standards as quickly as possible. Indeed, the rapid development of AI technology and regulation has created a new demand that organisations are struggling to meet. As international stakeholders are calling for more convergence on standards, pressure is increasing to adopt existing frameworks for AI. Furthermore, the field of expertise of ESOs, made up mainly of industrial experts, is not that of fundamental rights, and they could face a legitimacy problem if they tried to take on this role reserved for legislators and judges. However, if the ESOs fail to develop standards for AI that provide a sufficient level of protection against the dangers of these technologies, the AI Act risks becoming virtually unenforceable.

To counter this, the Commission has planned for a fallback solution: common specifications that could replace hENs for the AI Act in case standards are late or insufficiently address fundamental rights. However, the anticipation of such an emergency exit could also reveal a lack of confidence in the proper functioning of the European standardisation system, as the Commission might not trust the ESOs to deliver standards in line with its vision.

If AI standards are to deliver on their promise to safeguard fundamental rights in accordance with the AI Act, they will need to withstand international and industry pressure, while meeting deadlines. Time is of the essence, as the Commission closely scrutinises their progress.

Chapter 8
Behind the scenes of AI standards-making in Europe: interviews with CEN-CENELEC JTC 21 experts

Where does this chapter come from?

This chapter is an original contribution to the thesis. This study resulted from the collaboration with a PhD student in sociology at the French École des hautes études en sciences sociales (EHESS), Hélène Herman. It was born of the idea of combining our respective expertise in regulation and social sciences to analyse in greater depth the network of stakeholders working on AI standardisation. The interviews were conducted jointly, and we divided up the work of transcribing the interviews and collecting and labelling the data. However, I personally analysed the data and coded the interviews specifically for this chapter. The writing of this chapter is also mine alone.

This chapter is complementary to Chapter 7 in that it draws similar conclusions, but from fieldwork and interviews, whereas Chapter 7 focuses on a study of the literature.

This chapter was also released online as a standalone work, with a different title:

> Gornet, M. and Herman, H. (2024). A peek into European standards making for AI: between geopolitical and economic interests. `https://hal.science/hal-04784035`

Chapter's abstract

The new European regulation on artificial intelligence, also known as the AI Act, will require providers to assess their high-risk AI systems against certain requirements. To make this easier for them, a joint committee (JTC 21) of CEN and CENELEC, two European standards organisations, is preparing technical standards that will contain specifications on how to comply with the requirements of the AI Act. In this chapter we present the results of fieldwork within CEN-CENELEC JTC 21, where we attended group meetings, collected quantitative data on participants and conducted qualitative semi-structured interviews with 16 standardisation experts. The study aims to understand the day-to-day work of standardisation experts, who they are and what drove them to standardisation, as well as the difficulties they encounter. Our findings reveal that experts from different stakeholders groups – public and private, big or small companies and so on – do not have the same experience with standardisation. However, they agree that the standardisation system

is a complex machine, with processes that are difficult to comprehend, that there is a general lack of experts to work on AI standards, and that it is a highly diplomatic arena, with geopolitical and economic stakes. While many experts are highly critical of the way the current system works, their testimonies also point the way to potential improvements.

8.1 Introduction

Following the adoption of the European regulation on artificial intelligence, also known as the AI Act (European Parliament and Council, 2024b), the European Commission mandated two private standardisation organisations, the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC), to draft technical standards which could be used to support the essential requirements of the AI Act. CEN and CENELEC therefore decided in 2021 to launch a Joint Technical Committee on AI: CEN-CENELEC JTC 21. Experts in AI from both the public and private sectors, were to contribute to the work of JTC 21, including technical reports and standards, and specifically, harmonised standards (hENs) requested by the Commission.

The structure and *modus operandi* of standardisation organisations have long been of interest to the academic world. A number of criticisms are often raised, in particular the long-standing debate on the involvement of the industry in standardisation (Mattli and Buthe, 2003; Brunsson and Jacobsson, 2002; Büthe and Mattli, 2011) and the subsequent risks of conflicts of interest and policy capture (Mitnick, 2011; Whittaker, 2021)[1]. In Europe in particular, the standardisation system has been under scrutiny for some time. Indeed, under the New Legislative Framework (NLF) for product safety, hENs drawn up by European Standards Organisations (ESOs) have legal effects and a quasi-regulatory status[2]. For AI in particular, the AI Act will fall under this product safety framework and will therefore follow the same pattern of having hENs to specify the obligations set out in the text. Policy scholars have expressed concerns about the content of these standards, as well as the tight schedule under which the standardisation organisations are working (Perarnaud, 2023; Pouget, 2023). Some notably highlight the challenge of using the technical tool of standards to address a wide range of non-technical issues, such as societal or fundamental rights issues (Almada and Petit, 2023; Tartaro, 2023; Gornet and Maxwell, 2024)[3].

To the best of our knowledge, studies on standardisation are conducted from the outside, looking at the relevant literature to decipher the role of standards in the AI Act, their potential to solve AI-related issues or the risk of having technical standards to answer normative questions. Some reports list current standardisation initiatives, but they come from the Commission and not from independent academic research (Nativi and De Nigris, 2021; Soler Garrido et al., 2023a). The closest to our work might be the white paper on standardisation published by ZVKI (Baeva et al., 2023), a group of researchers from academia, think tanks and research institutes. In this white paper, the authors conducted an expert consultation to understand the role and pitfalls of standardisation for AI systems. However, this work is not intended to be a scientific study that maps out the space of AI standardisation, but rather a summary of current issues which aim to bring more attention to AI standardisation. For instance, not all the people interviewed are actively involved in

[1] For more information, see Chapter 7, Section 7.7.

[2] For more information, see Chapter 7, Section 7.3.

[3] For more information on standards on fundamental rights, see Chapter 5.

standard making. Furthermore, the names of the people interviewed are not made anonymous, which runs the risk of distorting their discourse and not revealing all the truths standardisation can entail.

In this chapter, we aim to go deeper than these previous works and go directly ask the people who work on AI standards – the so-called "standardisation experts"[4], to tell us about their work. We seek to decipher the world of standardisation, in particular the purpose of the experts' work and the dynamics within the system, as well as the specific features, if any, of AI standardisation. We notably seek to answer three research questions (RQs):

- **RQ1**: What is the general organisation of standardisation work within CEN-CENELEC JTC 21?
- **RQ2**: Do AI standards differ from other standards and how?
- **RQ3**: Do all experts have the same experience of standardisation and what externalities shape their experience?

We first present our methodology in Section 8.3, based on fieldwork and interviews with 16 standardisation experts, supplemented by the analysis of quantitative data on participation within CEN-CENELEC JTC 21. Section 8.2 provides the necessary background to understand the stakes of this work. In particular, it highlights the need for standardisation in Europe, as standards are intended to support legislation, such as the AI Act. It further provides a map of the standardisation ecosystem, including the organisations and working groups involved. AI standards are still in the making and represent important issue for the proper implementation of the AI Act. The European standardisation groups therefore appear to be a political arena, at the heart of regulatory development, where tensions between stakeholders are at their highest. Section 8.4 presents the general organisation of standardisation work. It allows us to determine who is involved in standardisation work and why, as well as how they work.

In Section 8.5, we analyse the originality of AI standards, which lies both in the complexity and novelty of the technology and in the way it is approached by standardisation and the political world that oversees its efforts. In particular, we show the uniqueness of CEN-CENELEC's approach to creating a single horizontal standard for AI compliance. For the first time, AI standards will have to deal with ethics and fundamental rights, which attracts new stakeholders to standardisation. Given the high stakes involved, AI standards are being closely monitored by the European institutions.

In Section 8.6, we analyse the plurality of discourses that coexist within standardisation, highlighting the common experience of standardisation experts, as well as their differences. We invite them to tell us about the difficulties they encounter in developing standards and their frustrations with the way the current standards system works. In particular, many of them mention the complexity of the standards development process, the current shortage of experts, the influence of non-European actors and the over-representation of industrial stakeholders. However, their discourse is shaped by the institution they represent, their background and their previous experience of standardisation. Some of their criticisms point to ways in which the current system can be improved or restructured. However, the changes seem difficult to achieve and not everyone agrees with them.

[4] In this chapter, we will refer to the people we interviewed and those who are involved in AI standard making as "standardisation experts", as this is the name generally used in this sphere, by SDOs and by the individuals working on standards themselves, to describe them (ISO, n.d.i). For more justification on this choice, see the box in Chapter 4, Section 4.5.2. To understand why the use of this nomenclature can be challenging, see Section 8.6.2.

Finally, in Section 8.7, we reflect on the reasons for the gap between the high stakes of standardisation and the low level of public interest in this issue.

As this work provides a few keys to understanding standardisation work, it can first be used as an educational document aimed at the general public. It can serve to inform them about the world of standardisation, to shed light on these important discussions which often take place behind closed doors, and to raise awareness on both the need for standardisation and the shortcomings of the current system. Additionally, we believe that this work can help those wishing to become involved in standardisation to get started and to become aware of the difficulties they may face. Finally, it can also help current experts to reflect on the difficulties of standardisation work and to open their eyes to the problems encountered by some of their colleagues.

8.2 Background

8.2.1 Why are standards important?

Standards are documents containing technical requirements or guidance, addressed to professionals, which codifies industrial expertise and compliance with which is voluntary[5]. According to this definition, standards can be developed by anyone, but certain standards have a special status in the EU. In particular, the EU distinguishes between mere "technical specifications" and standards drawn up by "recognised organisations", called Standards Developing Organisations (SDOs). There are six SDOs, and three which are located in Europe and are therefore called European Standardisation Organisations (ESOs). The three international SDOs are ISO, IEC and ITU[6], and the three ESOs are CEN, CENELEC and ETSI[7]. These SDOs are private bodies which work on standardisation topics as they see fit.

Standards also play an important role in Europe as part of the New Legislative Framework (NLF) for product safety. Products covered by European directives and regulations under the NLF must comply with a number of requirements defined in the legal text, but these legal requirements are supplemented by technical requirements defined in standards. Indeed, under the NLF, the European Commission can send a Standardisation Request (SR) to one or several ESOs, to ask them to work on specific topics for standards which will support European legislation, standards known as "harmonised standards" (hENs). The ESOs can choose to accept or reject the proposal, but if they accept it, they must present their work on these topics to the Commission within a deadline set in the request. In addition, the Commission has a supervisory role in the development of these standards as it is the final approver of hENs and can choose to reject the work of ESOs if it does not meet the requirements of the request or of the harmonised legislation it aims to cover (art. 10.6 Reg. 1025/2012[8]). The European Commission is therefore responsible for political choices while the ESOs are responsible for technical choices[9].

Furthermore, the Commission may choose to publish a reference to the hENs in the Official Journal of the European Union (OJEU), giving them special legislative powers, such as the power to grant a presumption of conformity with the corresponding harmonised legislation for stakeholders who comply with these standards. However, hENs are not necessarily developed by the ESOs themselves, since ESOs may choose to adopt international standards and submit them to the Commission for harmonisation. The standards developed by the SDOs therefore play an

[5] For more information on standards and to see where this definition comes from, see Chapter 4, Section 4.2.

[6] Respectively, the International Organization for Standardization, the International Electrotechnical Commission and the International Telecommunication Union.

[7] Respectively, the European Committee for Standardisation, the European Committee for Electrotechnical Standardisation, and the European Telecommunications Standards Institute.

[8] Regulation (EU) 1025/2012 (European Parliament and Council, 2012) on standardisation sets out the rules for the European standardisation system. For more information about the NLF and the different types of standards in Europe, see Chapter 4, Section 4.6.

[9] For more information on the role of the Commission and the delegation of power in standardisation, see Chapter 7, Section 7.4.

important role, as they can become hENs which are the main means of complying with European legislation. Moreover, even when they are not harmonised, standards have a strong influence on the technologies that can be adopted in the long term, which also gives them an economic role.

8.2.2 The ecosystem of AI standardisation

As part of the regulation of artificial intelligence, the European institutions have reached agreement on a legislative text which has just been published in the OJEU: the AI Act (European Parliament and Council, 2024b). The AI Act is part of the NLF, which means that for certain products covered by the AI Act, the text will define requirements that will be supplemented by technical requirements in hENs. This is notably the case for systems considered to be "high risk". These high-risk AI systems include products already covered by harmonised legislation, such as medical devices or machinery, as well as new applications such as biometric AI systems or systems intended to be used by law enforcement, border control or the justice system[10]. In order to provide high-risk AI systems with a means of complying with the requirements of the AI Act, the European Commission has submitted a standardisation request to CEN and CENELEC. The public version of this request was made public in the form of a Commission implementing decision in May 2023 (European Commission, 2023a). In the request, the Commission lists ten items that should be addressed by one or several hENs, corresponding directly to the requirements of the AI Act for high-risk AI systems[11]. It is not yet known whether another version of the request, or even a completely new request, will be issued to cover more topics. Topics that are not currently in the request, but could be in the future, include requirements for General Purpose AI (GPAI) systems, systems based on AI models trained on large amounts of data and which can be used in a variety of downstream tasks.

All six of the SDOs are currently working on developing AI standards, but the most important initiatives are being led by a Joint Working Group (JTC) between CEN and CENELEC – the CEN-CENELEC JTC 21 – and a Sub-Committee (SC) in a JTC shared by ISO and IEC – the ISO-IEC JTC 1/ SC 42. Indeed, JTC 21 is directly drafting standards to answer the request from the Commission, as well as other topics at their own discretion. But JTC 21 could also adopt ISO/IEC standards which are more advanced, either just to consider them as European standards, but also eventually to present for harmonisation. Indeed, mutual adoption of standards is made possible thanks to the Vienna agreements between CEN and ISO, which also account for possible parallel development of standards between both organisations. However, ISO always has priority over CEN to develop standards, if there are no specific reasons to leave the development to CEN[12].

They are several working groups (WG) tasked with different projects, both at ISO/IEC JTC 1/ SC 42 and at CEN-CENELEC JTC 21[13]. Some of these working groups are developing one or several standards on AI, such as WG 4, which is developing the future "AI Trustworthiness framework", a standard which should be presented for harmonisation, in response to the Commission's request[14]. To take part in the discussions of these working groups, an expert must first register in a National Standardisation Body (NSB), such as AFNOR in France. Like SDOs, NSBs are private entities and represent their country in standardisation discussions at European or international level. To join, experts must usually pay for annual

[10] To find out more about systems that are considered high risk, see Chapter 3, Section 3.5.2.

[11] For more information on the request, see Chapter 4, Section 4.7.3.

[12] For more information on the Vienna agreements and the problems of sovereignty it poses, see Chapter 7, Section 7.5. To see our experts' opinion on these agreements, see Section 8.6.3.

[13] To see the structure of CEN-CENELEC JTC 21 and ISO/IEC SC 42 and their relative working groups, see Chapter 5, Section 5.3.1.

[14] For more information on the AI trustworthiness framework, see Section 8.5.3.

membership fees, paid by their home institution. However, academics and small business experts are exempt from these fees. In addition, some experts from "partner" organisations can participate in CEN-CENELEC JTC 21 without registering with a NSB[15]. However, they cannot participate at international level[16].

Standardisation work is voluntary, which means that experts are not paid by the NSB or SDOs to work on standards. When they register, they agree to write standards anonymously, thereby conferring intellectual property of their work on the NSB. These NSBs then publish European and international standards and sell them to interested stakeholders.

8.3 Methodology

8.3.1 Data collection: participation to CEN-CENELEC

For this study, we registered with the French National Standards Body (NSB), the Agence Française de Normalisation (AFNOR), and asked to participate to standardisation initiatives at European level, particularly in the working group (WG) 4 on "Foundational and societal aspects".

As members of JTC 21 WG 4, we had access to the list of members who registered to follow the activities of this working group. They do not necessarily participate actively in JTC 21 discussions, but they are at least monitoring activities. For each of the individual who joins WG 4, CEN-CENELEC collects their name, e-mail contact and the National Standardisation Body (NSB) they are registered in. We used this information to build our own database, looking for information on the internet about the person's current professional status. Our database therefore contains the following information, for each member:

- "Id number": integer from 1 to 218, representing the expert from WG 4. We have not retained any names or contacts from the CEN-CENELEC database; everything is anonymised.
- "NSB": the name of the country in which the expert is registered.
- "Affiliation large": the type of affiliation between "Industry", "Consulting", "Organisation", "Research", "Government", "Standards".
- "Affiliation precise": a refined version of the affiliation types[17].
- "Affiliation origin": the country of origin of the main affiliation.

The labels were chosen by the authors in order to map the various sets of interests. All the data was subsequently annotated by hand by the authors. It should be noted that CEN-CENELEC also uses its own classification of stakeholders, but many data instances were missing and the categories were too broad to allow precise analysis. For example, all types of industry were grouped together. However, in our opinion, a large IT company ("BigTech"), a multi-national company applying AI in its sector ("Corporation"), an SME developing small AI systems and a consulting firm selling its expertise and services, may have different interests in AI standardisation and therefore deserve their own categories. The many categories we have created show the diversity of AI standardisation experts.

The data we collected shows that 218 people are registered at WG 4 alone. All of these people do not necessarily take an active part in standard drafting[18]. Conse-

[15] For more information on membership status in SDOs, see Chapter 4, Section 4.5.2.

[16] This poses issues when international standards are considered for adoption by European entities. For more information, see Chapter 7, Section 7.5.3

[17] For "Industry" and "Consulting", the refined categories are "Freelance", "SME", "Corporation" and "BigTech", representing the size of the company, with "Corporation" being trans-national companies outside of BigTech companies. "Organisation", are separated between "Professionals", "Thinktank" and "Consumers", with the latter encompassing all partner organisations. "Research" is separated between "Academia" and "Institute", with the latter representing private research entities. "Standards" represent standardisation organisations, both at national level with "NSB" and European level with "ESO". "Other" represent standardisation organisation that are not part of the official NSBs but are strongly related. Finally, "Government" is separated between "National", "European" and "International" organisations.

[18] See Section 8.6.2.

quently, analysis of this data does not directly show who contributes to standards. Rather, it shows who are the interested parties who follow the discussions on standardisation.

8.3.2 *Fieldwork and experts interviews*

For this study, we further interviewed 16 people from CEN-CENELEC JTC 21[19]. At the same time, we registered with the French National Standards Body (NSB), the Agence Française de Normalisation (AFNOR), and took part in standardisation discussions at CEN-CENELEC JTC 21, particularly in the working group (WG) 4 on "Foundational and societal aspects"[20]. There, we attended meetings to understand the dynamics underway within the organisation[21]. WG 4 is the group in charge of developing the "AI Trustworthiness framework", the main standard that will be used for compliance to the AI Act[22]. It is therefore the place where the stakes are highest, where many stakeholders seek to get involved, and where tensions between actors with competing interests are more likely to emerge.

For the interviews, we selected a sample of standardisation experts representative of the different types of profile that can be found in the field of standardisation, which we identified thanks to the analysis of the data collected from CEN-CENELEC[23]. We notably spoke to academic researchers, consumer representatives, people working in research institutes, private organisations, Small and Medium-sized Enterprises (SMEs), BigTech companies[24] and public authorities. The distribution of interviewees by stakeholder group is shown in Appendix A.5. Where interviews were conducted in another language and experts are quoted in this work, the translation was done by the authors and is indicated by an asterisk (*).

The interviews we conducted were semi-structured, with various themes to be explored. Our grid of themes can be found in Appendix A.4. Each theme was addressed in all the interviews, but the questions asked depended on how the interview unfolded. The coverage of certain themes and questions evolved as we interviewed experts, in order to obtain more precise answers and deepen our understanding of certain subjects. The interviews were conducted jointly by the two authors[25], combining legal and sociological expertise[26]. Interviews lasted between thirty minutes and three hours and were recorded with the consent of the interviewee. Only one interviewee did not consent to being fully recorded and asked for the recording to be interrupted during the interview. The recordings were translated from speech to text by the authors, with the help of a local instance of a machine learning model for speech recognition. They were then coded into themes that evolved as the interviews with the experts progressed. The following coded themes correspond to different sections of this chapter.

8.3.3 *Data collection: interviewed experts*

Experts were assigned an identification number from P1 to P16 to ensure anonymity. For each expert, we noted their main affiliation type similarly to the data collected through CEN-CENELEC[27] In addition, we labelled each experts according to their background: "computer science" – for experts in computer science and AI; "governance" – for experts in standardisation processes; or 'humanities' – for experts in other disciplines, such as social science, philosophy or law, or for representatives of interest groups such as consumer groups or trade unions.

[19] Interviews took place between April and October 2024.

[20] The data we collected also comes from WG 4, but the experts we interviewed came from the whole of CEN-CENELEC JTC 21.

[21] We have been part of AFNOR since January 2023. We started by attending AFNOR meetings, then we realised that most of the work was taking place within JTC 21 and we started attending WG 4 meetings in June 2023.

[22] See Section 8.5.3.

[23] See previous Section.

[24] BigTech companies are the largest IT companies in the world. They include American companies such as the GAFAM – Google, Amazon, Facebook (Meta), Apple and Microsoft – as well as IBM, Nvidia or Tesla, for instance. They also include Chinese companies such as the BATX – Baidu, Alibaba, Tencent, and Xiaomi – as well as Huawei, DiDi, or DJI.

[25] With the exception of two interviews which had to be conducted by a single author due to scheduling constraints.

[26] See "Where does this chapter comes from?" at the beginning of the Chapter.

[27] See Section 8.3.1.

We chose these categories because we found during the interviews that the experts themselves tended to distinguish between "AI experts", who have expertise in the object of standardisation, and "experts in standardisation", who have experience in the standardisation process and know better than AI experts how to develop a standard. We added a third category because we found it difficult to classify our experts into these two groups alone. This last category of experts seems to be quite unique to AI standardisation[28]. This third category allows us to bring together experts who are not generally found – or not very often – in other standardisation circles.

Consequently, we created two databases: one for CENCENELEC participation, described in Section 8.3.1, and one specific to the experts we interviewed, with an additional label on the area of expertise, which was too difficult to infer for the participants we did not interview. The data collected for participation in CEN-CENELEC and for interviewed experts is analysed in particular in Section 8.4.1. The rest of the sections are mainly based on the experts interviews. We have chosen not to disclose the CEN-CENELEC database, but the database specific to the experts we interviewed, with the distribution of their backgrounds and affiliations can be found in Appendix A.5.

8.4 *General organisation of standardisation work*

To begin with, we seek to understand the organisation of standardisation work within CEN-CENELEC JTC 21: who are the experts, what drove them to standardisation and how do they work.

8.4.1 *Meet the experts!*

To understand the dynamics at work within standardisation, we first need to look at who is working in this field.

A variety of paths to standardisation

Among the experts we interviewed, we noted that there are no two identical ways of getting into standardisation. Some experts are asked directly by their companies to work on standards [P3, P7], others are interested in the ethical issues of AI and come across standardisation [P2, P4, P10], others discovered standardisation with the AI Act [P5], and so on. It almost feels like some of them arrived here by chance, that they were assigned this task of monitoring or participating to standardisation work by their organisation, but were not specifically trained for this and had to discover this world for themselves.

Many experts have also entered this world thanks to someone on the inside who suggested that standardisation work might be of interest to them [P2, P3, P7, P9]. Indeed, it is not uncommon for JTC 21 members to recruit new members, and experts admit that they often try to bring in people with whom they have common interests. *"I had to bring fifteen [people] or so into CEN-CENELEC. And I am quite satisfied"**, says [P2]. This is an excellent way of increasing one's number of allies by specifically selecting people who have a similar vision and position to them. *"There are very few of us. So, everyone who seems to be interested, of course we want to bring them in. This will give us more weight when it comes to taking a stand and negotiating"**, says [P9]. In standardisation, everything stems from your network.

[28] For more information, see Section 8.5.2.

A highly interdisciplinary world

Standardisation experts have various backgrounds and experiences. Some experts are engineers, computer scientists, and got involved in standardisation because they were already technical experts the standardisation object – here, in AI. But just as many people appear to occupy governance and management positions and have acquired expertise directly in the field of standardisation. This type of experts are often involved in several standardisation projects at the same time, and not solely on AI. They are sufficiently competent in the various fields to understand what the computer science experts are saying. However, their role is not to take part in the technical discussions, but to provide support by helping with drafting processes and facilitate the development of standards. Yet, nothing provides them from giving advice on technical points too. This separation between "computer science" experts and experts specialised in standardisation itself, whom we called "governance" experts, is sometimes made by the experts themselves [P8].

But expertise can also vary according to the background of the experts. While a large proportion of experts in AI standardisation appear to be from computer science, some delegations, like the French one, are very varied and include a variety of profiles, such as ethicists, or legal scientists. AI standardisation is also particularly attractive to many civil society players who do not fall into this binary classification. We therefore added a third category, to represent these new stakeholders: the "humanities" experts. In the 16 experts we interviewed, we kept the balance between these three categories of stakeholders[29]. This is, however, not necessarily representative of all the experts who follow standardisation work. But we did not collect this information for our dataset of members, because we did not want to infer their expertise without discussing it with them.

Similarly, we did not collect the gender of all the members of WG4 because we did not have the means to deduce this information without talking to the individuals. However, it should be noted that several experts we interviewed explained that men are still over-represented compared to women in standardisation today [P2, P11]. In the experts we interviewed, three of them (18.75%) are women.

A large representation of stakeholders

The data on CEN-CENELEC members shows that behind the classic separation between public and private actors lies a wide variety of stakeholders. For example, in the private sector, the raison d'être of companies and organisations can be very diverse. Some companies develop AI systems, others are subcontractors, others implement AI in downstream tasks, and so on. In particular, we have chosen to separate traditional companies, which sell products, from consulting firms, which provide services and expertise. We have also separated transnational corporations from BigTech companies, because today, the GAFAM and BATX have far more power in the digital technology market than traditional companies. Private players also include private organisations such as professional and expert organisations, think tanks and a number of private research institutes. Organisations also include consumer and worker representatives from a handful of entities that have been selected by CEN-CENELEC to participate as partner organisations. Finally, members also include representatives of standardisation organisations, mainly the NSBs, but also people from ETSI who come to see how CEN-CENELEC's work is going, or

[29] 6 experts in computer science, 5 in governance and 5 in humanities. This distribution is available in Appendix A.5

other standards bodies that are not official NSBs. On the other hand, public actors include academics, national government bodies, the European Commission and a number of international organisations. The data is presented in Figure 8.1.

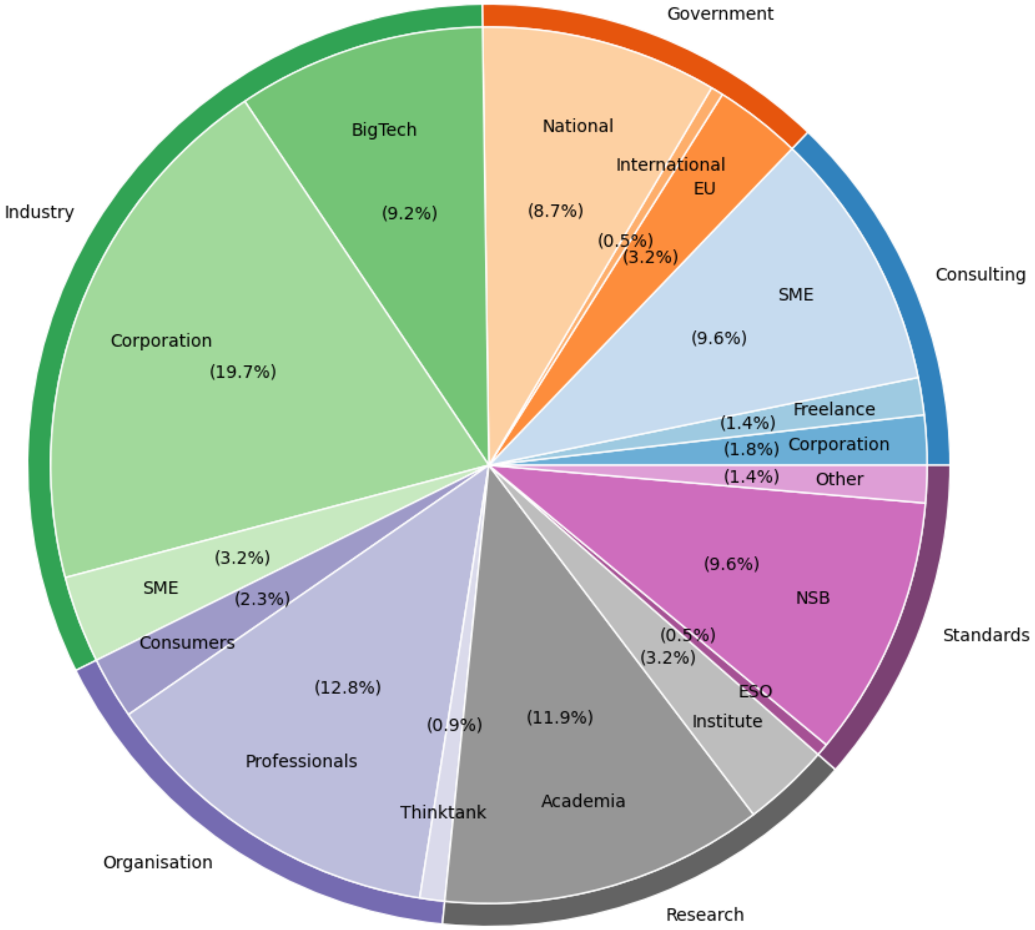


Figure 8.1: Pie chart of the main affiliation of CEN-CENELEC JTC 21 WG 4 members.

Firstly, we confirm a fact already widely known in the literature, namely that there is a large proportion of private stakeholders. If we consider that “public” actors are those labelled “Government” plus the academia, and that the rest are private entities, three out of four members are from the private sector. Company size is also important, with transnational companies, including BigTech and other corporations, accounting for almost a third of members.

It should be noted that we have displayed the main affiliation of the stakeholders, but many members have several affiliations. Some civil society representatives may also be part of a private organisation, work in a company or teach at university. Similarly, academics sometimes work for non-profits or are part of governmental bodies. Finally, some have changed position and are no longer affiliated to their former institution, but remain very close to its interests and concerns. It is therefore very difficult to have a complete representation of reality.

Some countries are more involved than others

In our dataset of WG4 members, we also looked at the country of origin of the experts and their home institution. Once again, it should be remembered that these figures only represent those who follow the discussions, and not necessarily those who actively participate in them. Yet, they do give an idea of the power dynamics between the countries involved. The data is presented in Figure 8.2.

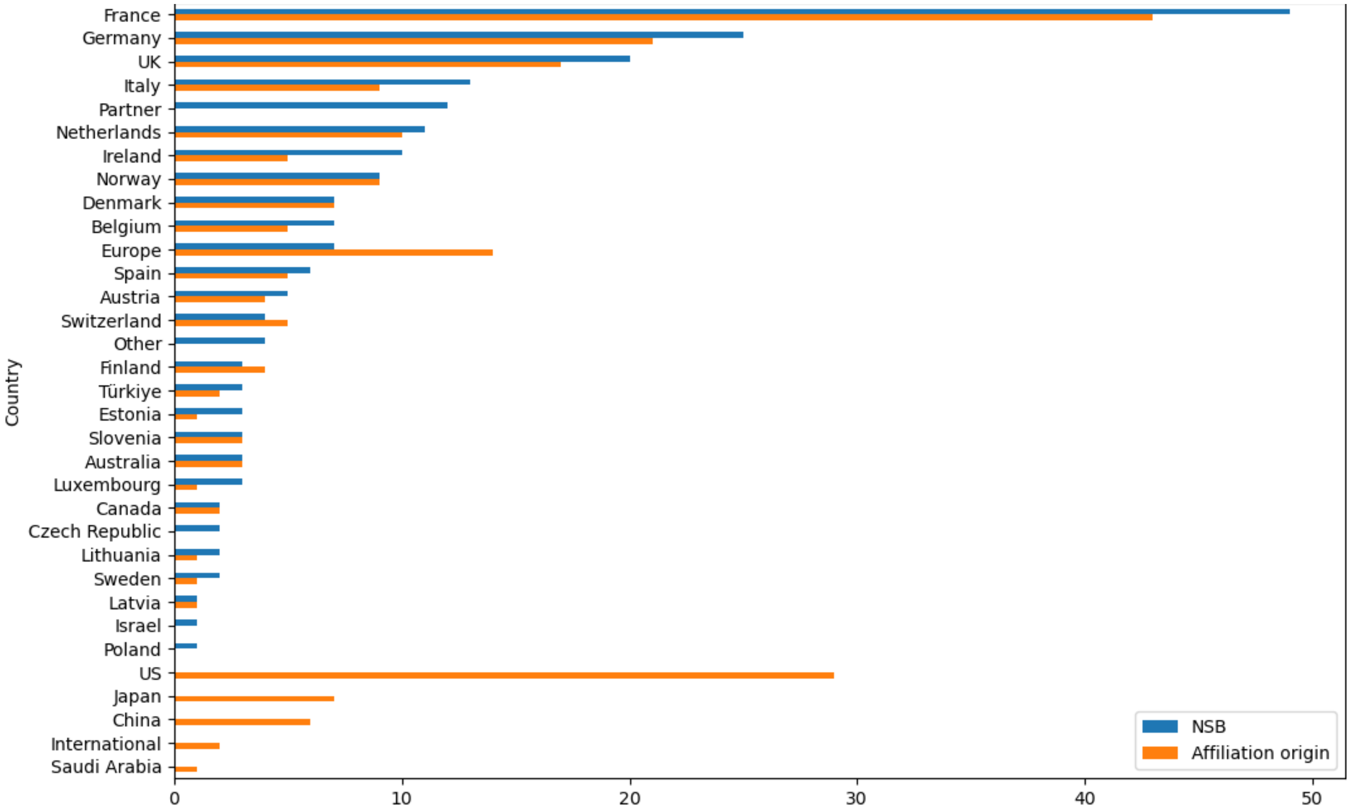


Figure 8.2: Bar chart of WG 4 experts' NSB country of origin and country of origin of their affiliation.

First of all, there is a major imbalance in the number of experts per country. France is well represented, with 49 experts out of 218, followed by Germany, the UK and Italy with 25, 20 and 13 experts respectively. Behind Italy, partner organisations such as consumers and workers representatives, which are not affiliated to any NSB, have 12 members. We also note the presence of countries outside the European continent, such as Canada and Israel, which are authorised by CEN-CENELEC to participate but do not have voting rights. In addition, certain countries located in Europe, but not part of the European Union, have a full membership status, which means that they can participate and vote on standards in the same way as other members. The case of the UK is perhaps the most interesting, as the British delegation is strongly represented and exerts a certain influence on European standardisation[30]. While for some of the experts we interviewed, this is normal, since they will be affected by the AI Act and its standards [P9], for others, it calls into question the EU's sovereignty over its own standardisation system [P5].

Furthermore, Figure 8.2 shows that the nationality of companies and organisations does not necessarily follow the same pattern as the experts' countries. Indeed, many trans-national groups with a subsidiary in the EU are authorised to participate, even if the company's head office is outside Europe. This is particularly true of American BigTech companies, as well as certain companies based in Japan and China.

[30] It should be noted that other countries enjoy this status at CEN-CENELEC without being part of the EU, such as Türkyie, Norway and Switzerland, as well as other countries not present in WG 4. For more information on the different membership statuses at CEN-CENELEC, see Chapter 4, Section 4.5.2.

8.4.2 *The benefits of working in standardisation*

Standardisation work does not attract many people as it is voluntary and time consuming. Yet, many experts have found ways to make this investment of time pay off.

Getting an edge

In Europe, harmonised standards are a sub-set of the law. Working on standards allows experts to have a direct say on the content of the frameworks that will ultimately underpin the law. Ultimately, it is a means of influencing the law itself. Even when standards are not harmonised, when they are developed by international organisations such as ISO, they are often very well received and adopted worldwide by companies. Thus, even outside of legal influence, stakeholders are seeking to seize the economic power of standards.

According to [P2]: *"It is a prospective role. [...] If you can put the right words in the right places, you can do good things"**. On the contrary, if you do not participate in standards development, you risk falling behind. In the words of [P11]: *"When it comes to standardisation, you're either on the menu or at the table. [...] If you do not take part in the work, the subject will be dealt with without you and you will suffer the consequences"**.

Additionally, working in standardisation gives experts practical knowledge about of what is on the horizon, whether in terms of regulation or scientific innovation. It enables industry stakeholders to stay one step ahead of new technologies and can help decipher geopolitical issues otherwise reserved for diplomats, as well as providing a better understanding of issues of sovereignty and competition [P11]. Standardisation also provide experts with a large network of stakeholders. It enables them to meet new people interested in similar areas of innovation, and then to build projects outside standardisation [P8]. This is particularly true of companies and research institutes, which can collaborate on industrial projects.

Proposing products that align with standards

Most of all, companies have strong business incentives to develop products that comply with current standards, so that they can be more easily implemented with their partners or customers [P9]. Some experts explain that they first became interested in a standard because they were planning to develop a product and wanted to see what the standard provided before starting development [P3]. Hence, because they have been involved in the creation of the standard, companies can develop a product that fully meets the expectations of the standard. Having a product that is already aligned with standards also gives companies a competitive edge on the market [P9]. It is a marketing argument, as buyers will prefer to buy a product which already complies with standards and regulations.

However, sometimes, participation in standardisation is not motivated by the development of a product based on the standard, but rather by the development of a standard based on a product. Even if standards never directly impose the use of a specific product, the requirements they contain can be oriented in such a way that the product under development meets the need formulated in the standard [P3]. Companies can then advertise their product as a solution that complies with the standard. [P9] tells us: *"It is self-serving. The day I want to do business in this country,* *I will benefit from a more favourable environment because I will speak my own language. '[This ML concept] is define in this document you're using. It's a good thing my product does exactly that.' So for me, this is indirect business development. It is in my interest to get as many states as possible to align with my vision of things, because that will help my business"**.

Creating network and selling expertise

Standardisation is a great space to grow one's network. Experts from different sectors explain that it is a way for their company or institution to be seen and build partnerships with other actors, outside of standardisation [P15]. Some experts are well established and well-known in the world of standardisation. They have begun to build their reputation and sell their company's expertise. They report that their reputation spreads by word of mouth within the standardisation ecosystem, enabling them to take part in more projects and win new customers [P9]. This is particularly true for previously unexplored subjects where these experts are the first to plant their flag and impose their vision, which has enabled them to be recognised today as the international experts in the field [P9]. It also enables experts from small businesses to gain an edge over large companies. [P9], who runs an SME, explains: *"Once we started to get to grips with the subject and people saw that we were the ones holding the reins, the balance of power changed between us and the big groups. [...] We say to companies: 'We are writing the standards that will regulate you. [...] If you want, we can now help you to comply with it'"**.

But even outside the standardisation ecosystem, experts sell their standardisation expertise, often as consultants [P2]. For [P8]: *"Understanding what a standard is and how it is made really helps. [...] Ultimately, it leaves a mark on my job, on my professional activity"**. It is a way of reusing information [P2] and compensating for the time investment [P11]. There is a large number of consulting firms whose business model is based on advising on standards and supporting companies. The consultants explain to stakeholders outside the ecosystem what standardisation is and how to comply with the standards in force [P4]. This consultancy activity can be carried out in parallel with the involvement of an expert in standardisation activities, but it can also be a springboard for a career after standardisation [P11]. Some are even using this expertise to create a quality label for AI or a private certification scheme [P8]. These schemes are not intended to compete with standards developed by recognised SDOs, but to provide a quicker and simpler way of showing consumers that a company is concerned about producing quality products. However, not everyone is in favour of these quality labels, as some experts strongly criticise them and consider them to be scams [P9].

[P11] summarises the benefits of getting involved in standardisation work: *"You can expand your network, you can sell your product, you can get to know people, you can advertise on LinkedIn about the various projects you have set up, you can have a name as a contributor or editor on a standard. But on the other hand, it requires a considerable effort. So you have to weigh up the pros and cons carefully"**.

8.4.3 A multi-national system based on consensus

The SDOs are complex structures, each with its own *modus operandi*. However, the roles taken on by the experts and the types of meetings they attend are often similar from one organisation to another.

A structure based on national delegations

At CEN-CENELEC and ISO/IEC alike, the standardisation processes are centred around national delegations. Experts need to be registered in a national body to participate in standardisation work[31]. Once registered, they can ask the secretariat of their NSB to join European or international working groups, as experts see fit and according to their areas of interest.

The more technical discussions happen in working groups, with interested experts. Each working group is based on a general theme such as "cybersecurity" or "societal aspects" [32]. Working groups meetings usually happen online. In a working group, experts can make contributions, i.e. documents often representing their own work and which are intended to feed discussions and sparkle debates [P4, P6]. Each working group has a "convenor", a person responsible for coordinating discussions within the working group and designated among the experts. The role of the convenor is to provide the group with the space and resources it needs to have fruitful and effective discussions. When a draft standard is launched, each project is allocated to an "editor" – or several co-editors. The editor's end goal is to reach consensus on a standard and publish it. They coordinate the work of all the experts on this specific standard [P10], and are responsible for gathering and compiling comments and for directing the drafting process [P2]. One working group can thus host several draft standards.

Alongside the discussions in the working groups, the experts registered with a NSB also meet online, generally once a month [P10]. This NSB meeting serves as a means of coordination, to inform everyone of the discussions taking place in the various working groups in the different SDOs. This enables experts who are not part of all the working groups to keep up to date with the work of others. These meetings also enable experts from the same NSB to discuss the country's position for the forthcoming plenary meetings where all countries are represented. Coordination within the NSB enables all stakeholders to find common ground and present a united front. The meeting is also an opportunity to send comments on a given standard and to vote on the adoption of standards.

Some NSBs, such as the French one, also organise "task force" groups, outside of the official NSB meeting, to discuss relevant topics with interested experts. This facilitate the formation of a national position on the topic which can later be proposed by these task force experts at the NSB's meeting [P15].

Once every few months[33], experts meet in-person. These plenary sessions last for a few days[34] and are organised around heads of delegation, i.e. experts mandated by their NSB to represent their country's position [P8]. The head of delegation is usually accompanied by two or three other experts from the same NSB, although this varies according to the size of a country's NSB and its involvement in standardisation, as participation is always voluntary. The plenary session consists of an opening meeting and a closing meeting, between which working group meetings are held.

At the opening and closing meetings, discussions are animated by the Committee chair. The heads of delegation are the only experts entitled to speak directly, but they may give the floor to other members of their delegation to defend an argument [P9]. When decisions must be taken, they are reviewed at the plenary meetings, one by one. Everything is planned in advance: proposals for new projects are sent out

[31] There are only a few exceptions, such as certain European organisations and non-profits, which are authorised to participate directly without going through a NSB.

[32] To see an overview of the different working groups within CEN-CENELEC JTC 21 and ISO-IEC JTC 1/SC 42, see Chapter 5, Figure 5.1.

[33] Usually three months for CEN-CENELEC JTC 21 and six months for ISO-IEC SC 42.

[34] Usually three days for JTC 21 and five for SC 42.

a few days before the meeting, questions and comments are sent back and answers prepared. So the whole discussion is orchestrated [P8]. However, it often happens that debates emerge and the discussion moves away from what had been planned at the NSB meetings. In these moments, the heads of delegation have to improvise, while respecting their country's general position. [P9].

At the end of the session, the plenary body reviews and validates the actions that have been taken [P8]. For instance, to launch a standard, the subject is proposed by experts with the support of a country. The proposal is then presented at a plenary meeting by the head of delegation. There must be a consensus between all countries to launch the initiative [P9].

Consensus building

All the decisions taken within the SDOs are based on consensus. However, defining consensus is a complex issue. [P8] reckons that it is not like a vote with a clear rule. Various experts have given us different definitions which, taken together, begin to paint a picture of how consensus is understood in standardisation. [P2] defined it as a *"systemic acceptability"**, [P10] as a *"lack of a sustained opposition"**. For [P8], the idea is that not everyone will agree, but that those who do not agree will not veto. These descriptions enable us to identify two key elements of consensus within SDOs: a majority of people must agree and there must be no strong opposition. It should be noted that this definition is close to the general meaning of the word "consensus", which can mean, for example, *"general agreement"* or *"the judgment arrived at by most of those concerned"* (Merriam-Webster Dictionary, n.d.e). A key element added by standardisation experts is the absence of strong opposition, something the experts verified experimentally.

However, it remains to be seen what the terms "majority" and "strong opposition" mean. [P4] explains that a "good" majority is usually considered to be around 2/3. This means that around 70% of countries have to agree for a project to go ahead [P10]. However, this threshold is highly theoretical. It depends very much on the decision and its context. Consensus at plenary meetings, for example, is different from consensus within working groups. As regard to the the term "strong opposition", [P10] explains that opposition alone is not enough, that it must also be supported by good reasons which must be communicated. For [P4], it depends on whether you have a good argument to put against the decision. This form of decision-making, based on consensus, means that there is no real decision-maker, although certain positions within the SDOs[35] have a role to play in facilitating consensus [P10].

Finding consensus in a highly diverse world

In standardisation, experts come from everywhere, do not necessarily speak the same language[36], have a variety of backgrounds, affiliations and fields of expertise. This diversity in expert profiles sometimes makes it difficult to understand each other. Yet they all have to agree to reach a consensus.

This diversity causes first a problem of linguistics and semantics [P2]. Experts from different backgrounds tend to talk about the same thing in different words or different things with the same word, which sometimes leads to misunderstandings. *"The problem is that everyone arrives with their own terms, their own definitions,*

[35] Convenor, editor, head of delegation, chair, etc.

[36] Work within ESOs and international SDOs is carried out in English, but for most experts this is not their first language.

*their own concepts, their own metrics"**, [P4] explains. This results in a huge amount of moderation on the part of convenors and editors, to try to understand and accommodate everyone [P2]. [P3] concedes that the end result is often a patchwork of concepts picked up here and there. Some experts admit that they enjoy these interdisciplinary discussions [P10], while some others find them tedious and tiring. *"It can be day-to-day debates about a semicolon or a term"**, [P11] sighs. Yet this work of defining concepts and terms is a central topic in standardisation. *"80% of the work consists of knowing what we are talking about. It is all about terminology. Words are important. Once you know what you are talking about, setting technical requirements is much easier"**, explains [P4]. Indeed, words are important because each stakeholders has an interest in using one or the other, to steer standards in a certain direction, for instance to encourage the use of a certain product or a certain form of regulation. One of the main objectives of consensus is therefore to align everyone's interests. For [P8], this is the mission of standardisation as a whole: to capture the interests of all stakeholders in society, to ensure that everyone is represented and to converge towards a common solution. In order to respect everyone's opinion, the experts are encouraged to follow a code of conduct provided by the standardisation organisation (CEN-CENELEC, 2018), whose main points are reminded to participants before each meeting [P11].

A game of alliances and diplomacy

Whether it is the industry or civil society and academia, every experts in standardisation is playing a strategic game of alliances to achieve their goals. Stakeholders must form coalitions with other countries or other groups, in order to increase their numbers and get their proposals through or block those of others [P4, P11]. *"We often say that everything is decided at the coffee break. In plenary meetings, all we do is confirm the positions that have already been defined"**, says [P9]. For [P16], there is always a lot of *"exchange of favours"*, which sometimes makes it difficult to understand why a certain stakeholder acts in a certain way or why a project fails.

These alliances are not always easy to make. As [P12] explains, experts need to find people who are not totally aligned with their positions in general, but with whom they can find common ground on specific points. [P4] notes that this type of strategy is the speciality of BigTech companies, which, according to them, have a common interest in the absence of standards or in empty standards[37].

But other groups of stakeholders also admit to using these strategies. Small companies or small countries can form alliances against other companies, so that when the latter tries to block a project, they have sufficient numbers to fight back [P9]. Civil society experts make alliances between organisations. But they can also join force with academic researchers and SMEs [P6, P13]. *"We collude sometimes, kind of like, 'what are we going to do? How are we going to do this?' And kind of share strategy. It is not a formal shared strategy because, you know, we want to maintain a kind of independence"*, recounts [P7].

However, it should be noted that these alliances are not fixed in time and can evolve rapidely. [P10] explains: *"Standardisation is a very dynamic environment and things can change very quickly. You might have the idea that there are kind of blocks, [...] but this is not always the case. Many times you can find people that are on different sides and they agree on something and disagree on something else. So the relationships are much more entangled"*. As there are many incentives to collaborate, [P15] notes that people

[37] See Section 8.6.4.

have no problem mixing between the public and private sectors. Standardisation therefore appears as a highly political arena, where relationships are central and diplomacy necessary.

8.5 The originality of AI standards

On the surface, AI is not that different from any other product that can be standardised. Several of the experts we interviewed had prior experience in standardisation, either in IT and software in general, or in cybersecurity. Some experts initially claimed that there was no real difference between AI and these previous works [P11]. But in reality, by digging deeper, we found other differences between the way standardisation used to work and the way it works today for AI.

8.5.1 A complex and new object to standardised

AI is a complex technology whose science and regulation are not yet very mature

AI, unlike other software that experts have worked on before, is a probabilistic system. This means that its outcome is not easy to predict. [P15] noted that there are not just one or two parameters, AI is very complex and its field of application is ill-defined. This means that the same system can act completely differently in two situations, such as an autonomous car in sunny or rainy weather. This complexity and unpredictability makes it difficult to standardise these systems in a general case.

In addition, the regulation of AI has only just begun. [P12] explains that, compared with data protection, for which principles have existed since the 1970s, efforts to "structure" AI are fairly recent[38]. This recent development is giving rise to lively debate about who should set the limits and what those limits should be. For [P9], the difference lies in the instability of this regulation. They explain that the rights enshrined in the GDPR are fixed, whereas for AI they are still under construction[39].

Furthermore, while AI itself is not a recent field, dating back to the 1950s, advances in machine learning are fairly recent and the technology is not well understood [P15]. For [P16], it is ultimately a question of whether there is already a consensus on the state of the art. For AI, the state of the art is constantly evolving, whereas for other sectors where standardisation works well, such as medical devices, the state of the art has been stable for years. [P3] recalls that when work began at JTC 21, there were still a lot of fantasies about what AI was. Generative AI did not even existed. The field of AI is evolving rapidly, and standardisation efforts are struggling to keep pace. Until 2018, there were no standards for AI. The initial work launched by ISO therefore had to start from a very high level and be very general. Today, [P3] acknowledges that there is a desire to define more precise requirements in the standards, but the field is not sufficiently advanced. *"[Standards on AI] are desperate attempts to control something that is beyond our control"**, [P3] says. There is a risk that technology will evolve faster than regulation and standards. *"It takes almost three years to develop a standard. [...] And when I look at the evolution of AI between 2021 and 2024, it is enormous. And I was wondering whether we would not have obsolete standards before they are even published"**, says [P15]. However, all the experts insist on the need for standardisation, even if this takes time. *"We can't just sit back and do nothing"**, concludes [P15].

[38] By way of comparison, the Organisation for Economic Co-operation and Development (OECD) AI Principles were first adopted in 2019, while the OECD Privacy Principles date back to 1980.

[39] It should be noted that at the time of this interview, the AI Act had not yet been published, but that the three European institutions had already agreed on the text.

The IT world is not used to drafting hEN

While hENs are not new in Europe and are already widely used for product safety[40], standardisation experts working in the IT sector are not used to developing such standards [P11]. Indeed, digital technologies have not been regulated through hENs, neither for the GDPR nor for any other previous piece of legislation. Therefore, experts do not know how the NLF works [P16]. [P5] explains that standardisation experts are not used to working with constraints. Sometimes, they want to say certain things but cannot because they have to stick to the framework of the AI Act and what is already defined in the legal text. But for hENs, experts cannot change the wording or obligations laid down in the AI Act, just complete it [P16]. *"We don't know how far we are allowed to hold the pen in this thing"**, says [P9].

Timing constraints

Most importantly, hENs come with a fixed deadline: the standardisation request sets the date for publication of hENs at 30 April 2025. This date cannot be postponed, as the hENs define the technical criteria that will be used by companies to assess the conformity of high-risk AI systems. Yet, this part of the legal text will come into force in summer 2026 and providers need time to adapt to new standards. For [P3], this deadline is a good thing: it acts as a driving force that pushes experts to work faster and results are obtained much more rapidly than in other fields. However, [P6] notes that this puts a lot of pressure on the experts to deliver results quickly, and increases their workload. [P12] warns that, because of the deadline, experts are working in a hurry, and that if they are rushed into developing standards quickly, this could undermine the quality of the final result. [P5] explains that, usually, it can take up to five years for a standard to be published, if at all.

JTC 21 began work on AI standards at the end of 2021, but at that time the ESOs were still waiting for a standardisation request. The first draft of the request was only sent to them at the end of 2022. [P8] believes that JTC 21 wasted a lot of time at the beginning. At first, the proposal for harmonised standards were more technical, on the evaluation of AI systems. Then the scope shifted, with a view to future harmonisation. But it has taken almost two years, since work began at CEN-CENELEC in early 2022, to define the right direction to take. Even if everything went according to plan, the publication of hENs at the beginning of 2025 already seemed optimistic. But given the delay in the launch of the trustworthiness framework, some experts believe that this deadline is no longer anything more than wishful thinking. [P11] admits that there is a growing discourse within JTC 21, putting forward excuses of a lack of material and human resources that would justify the delay of standards. Experts were reluctant to give us an estimate for the publication date, but a few unofficial dates circulate: initially postponed to August 2025, it seems now more reasonable to expect them around June 2026 [P5], although nothing was confirmed by the Commission nor the ESOs.

8.5.2 AI raises ethical and fundamental rights issues

The difficulty in standardising ethics and fundamental rights

AI raises ethical and fundamental rights questions, and the AI Act aims at protecting these fundamental rights. It is therefore clear that these normative questions

[40] For more information about the role of hENs in European law, see Chapter 4, Section 4.6.

will have to be tackled in standards[41], and initiatives are being launched on these topics.

Ethical and fundamental rights standards for AI are a new area of interest for standards committees, far removed from their usual work, and these new topics are not very well received by standardisation experts, who are used to more "traditional" engineering standards. [P11] explains: *"In the IT world, we believe that ethics cannot be standardised. It is not something tangible. [...] So, we are not rejecting these topics, but we are not betting too much on them either. We know we are going to do something on a fairly high-level framework, but we are not going to go into detail because that is impossible"**.

[P10] remembers that many experts were worried when JTC 21 first proposed ethical standards. [P10] explains that critics said that the topic was *"too normative"*, that standards should stay technical and that standards *"cannot decide what is good or wrong"*. But for [P10], this is a misunderstanding of the proposal. The purpose of standards relating to ethics is not to impose decisions, but to create an environment in which ethical decisions can be made [P2].

Similarly, legal experts are reticent about the idea of standardising fundamental rights. According to [P6], *"fundamental rights should not be standardised"*, only laws adopted through a democratic process can be. But what can be standardised are the conditions under which the use of AI complies with fundamental rights. For [P7], standards on fundamental rights should allow greater transparency and accountability for companies, in order to prove that companies have taken into account the risks to fundamental rights. However, [P6] admit that safety requirements defined by standards in other harmonised union legislation were easier to define than describing in a standard what has to be done to avoid being in breach of fundamental rights[42].

New interests in AI

Several experts noted that, unlike other IT areas, consumers and workers associations are heavily involved in AI standardisation [P2, P3, P10, P11, P16]. [P12] hypothesises that this is because they feel more concerned by AI and its dangers than by traditional software. For [P10], AI raises ethical issues that affect everyone, so civil society groups necessarily feel more legitimate to take part in the discussion. The hype of AI technologies is bringing more people in, because everybody wants to have a say on AI [P10, P16].

However, it should be noted that this diversity in AI standard setting is encouraged by the Commission. In both the AI Act and the standardisation request, the Commission calls for a "multi-stakeholder participation" to standardisation[43]. This include bringing more SMEs, more societal stakeholders such as consumer representatives, and more academics to standardisation. While this perspective is welcomed by most of our experts, some mention that the inclusion of civil society stakeholders, who are less competent in technical matters and strongly defend their interests, complicates the discussion with industry actors and consensus building, which slows down the development of standards [P16].

[41] For more information on how the AI Act tackles fundamental rights and invites to the standardisation of normative questions, see Chapter 5, Section 5.2.

[42] For a more complete discussion on how standards can address fundamental rights, see Chapter 5, Section 5.4, and Chapter 9, Section 9.3.2.

[43] For more information, see Chapter 7, Section 7.9.1.

8.5.3 High-level frameworks to tackle AI issues

Experts in CEN-CENELEC JTC 21 are currently working on several standards for AI[44], including standards intended to become hENs in support of the AI Act, but also other standards at the experts discretion. The field of AI is not yet very standardised [P12], but numerous initiatives are emerging[45], so much so that some experts speak of a *"deluge of standards"** [P4].

The AI Trustworthiness framework

The standardisation request to CEN-CENELEC includes ten subjects, but they will not necessarily map to ten standards. SR items can be addressed together in a standard, or separately in different standards, at the discretion of the ESOs. JTC 21 decided early on to group together all of the SR items in one single standard: the "AI trustworthiness framework".

The framework is intended to be very structural and an entry point in standardisation [P4]. It is intended to be *"the glue"* between every item in the standardisation request [P16]. Indeed, the ambition of the framework is to be an umbrella standard covering the entire requirements of the AI Act, and which could point to resources, such as other standards, to enable interested stakeholders to go into more detail [P9]. The intention was to make it easier, for small companies in particular, to comply with the AI Act, so that they do not have to read dozens of standards [P9]. But this effort to simplify has proved to be a complex task which involves a great deal of work and pressure. Coordinating the work is one of the first problems, as different groups are working on different parts of the standards to cover different requirements. All of these contributions are interdependent, but not all groups work at the same pace [P11]. This titanic task is unlike any other standardisation work, and experts report that they have never seen anything like it [P11]. Indeed, standardisation generally aims to avoid overlap, but the trustworthiness framework covers many concepts that are already covered by existing standards, notably at ISO[46] [P8]. This leads experts to wonder about what the final result will be like.

In particular, the question arises as to what level of granularity the standards' requirements should adopt. Experts have agreed to include in the standard what they call HLR, for High Level Requirements. [P9] explains: *"[HLRs] are fairly generic, but they still give you something to work with for implementation purposes. They are not precise enough to tell you exactly how high a threshold should be, or that you need to use this or that technique to validate it. But they prepare the ground and introduce good practices and ideas. Their level is therefore a sort of intermediary between the law that tells you 'you have to be like this', and ultra-technical stuff"**. However, problem arise when you consider each item independently. [P16] explains that for some requirements, there are already standards to draw on, such as ISO standards for data. But for others, such as human oversight, the AI Trustworthiness framework will be venturing into uncharted territory. For [P15], the AI Trustworthiness framework will end up being a mapping between the AI Act and other standards, referring for each item of the standardisation request to other more specific standards. However, as explained by [P16], it cannot refer to too many other standards, otherwise it would make compliance more complex, rather than simplify it. This simplification task is therefore enormous, as the Trustworthiness framework already contains hundreds of requirements, excluding the ones from outside standards [P16].

[44] For a complete list of JTC 21 published standards and standards under development, see (CEN-CENELEC, n.d.b,n).

[45] To see an overview of different initiatives, see Chapter 5, Section 5.3.

[46] Such as robustness or transparency.

Other standards

AI standardisation efforts at CEN-CENELEC are now focusing on the development of standards for the standardisation request [P10]. Even some experts who originally worked on ISO standards admit that they have given up their international work to concentrate all their energy on hENs [P3]. However, work at SC 42 is also still in progress and well ahead of that on JTC 21[47].

Another standard is being developed in parallel with the AI trustworthiness framework, with a view to being harmonised: the "AI risk management" standard. This particular standard posed some problems because another standard covering the same subject had already been published by international SDOs: the ISO/IEC 42001 (2023b) on AI management system. Thanks to the Vienna agreements, CEN-CENELEC is in the process of adopting the ISO/IEC standard at European level. However, the European Commission warned JTC 21 that the standard alone was not sufficient to cover the SR item on risk management and the corresponding essential requirement of the AI Act set out in Article 9 [P6]. At the time of the interviews, it was not yet known whether ISO/IEC 42001 would seek further harmonisation with substantial changes to adapt it to the European context, or whether JTC 21's standard on AI risk management would be the only hEN to address Article 9.

Apart from the two standards on trustworthiness and risk management, JTC 21 is also working on a variety of other standards, including standards to address ethics and fundamental rights questions. Examples include "Competence requirements for AI ethicists professionals" and "Impact assessment in the context of the EU Fundamental Rights" (CEN-CENELEC, n.d.d). However, as participation in the standardisation work for the two future hENs is very time consuming, experts are less inclined to take part in these groups [P10]. But the question of what will become of these draft standards is of great interest, as it will shed light on the possibility – or impossibility – of developing standards on ethics and fundamental rights questions[48].

8.5.4 The Commission supervises closely AI standards

The Commission supervises standardisation work

The European Commission is responsible for requesting and validating harmonised standards[49]. Even before the AI Act, the Commission already had a team responsible for overseeing standardisation work[50] [P4]. But with AI, experts report that the Commission is more involved than usual, intervening more and more in the standardisation process [P4, P6]. [P9] explains that the Commission's employees take part in the plenary meetings but much less in the working groups. Their objective is not to help draft standards, but to ensure their proper development.

This increased interest in the work of CEN-CENELEC can be explained by the high stakes around AI standards, as the AI Act strongly relies on them for practical implementation. For [P6], a representative of consumer interests, this is also due to the worrying state of the development of standards, which are lagging behind international standards and influenced by BigTech companies. According to [P6], the Commission is aware of the problems and is a little worried, but their participation is essential to keep the ship afloat. The Commission's main role is therefore to ensure that standardisation goes in the right direction. In particular, they check

[47] See all SC 42 standards on their website: (ISO, n.d.f).

[48] For more information on the difficulty to standardise ethics and fundamental rights, see Chapter 5, Section 5.4, as well as our discussion in Chapter 9, Section 9.3.2.

[49] For more information on the role of the Commission in standardisation, see Chapter 7, Section 7.4 and Section 7.9.

[50] For more information on the governance of the Commission with regards to standardisation, see Chapter 4, Section 4.6.5.

whether the standards that are supposed to support the AI Act are in line with the legal text [P4]. [P4] recalls that, initially, the industry was opposed to the AI trustworthiness framework, but that the Commission supported the initiative and helped it to see the light of day.

However, if Commission's employees are respected and listened to within JTC 21, they cannot be too active in the discussion [P5]. *"They don't want people to say that it is the European Commission's standard"**, [P4] reckons. They are therefore very careful to ensure that standards come from the industry, from field expertise. For [P4], the real problem is not that standards come from the industry, but that the industry represented is mainly made up of non-European companies that are trying to sabotage hENs. This supervisory role creates a complicated relationship between the JTC21 experts and those from the European Commission, the latter not wishing to get too involved but sometimes being forced to intervene and reiterate the objective of the work on hENs. At the same time, [P14] explains that *"they don't want to tell [JTC21 experts] what to do"* and are *"concerned about taking position"*. According to [P14], there is still work to be done to improve the way this relationship works, as the Commission is slowly figuring out its place and role.

The imperfect alternative of common specifications

The AI Act provides for an alternative solution in the event of the failure of hEN, called "common specifications". Common specifications are a type of technical standard[51], which can be established by the Commission in several cases: (i) when the standardisation request has not been accepted by the ESOs; (ii) when hENs are not delivered within the deadline; (iii) when the proposed hENs *"insufficiently address fundamental rights concerns"*; and (iv) when the hENs do not comply with the request (art. 41 AI Act)[52]. For [P6], this is a means put in place by the Commission to counter the ESOs' monopoly on drawing up standards and prevent the system from coming to a standstill.

Indeed, given the current state of the standards intended to be harmonised, with significant delays in the development process, it is likely that they will not be ready in time to meet the deadline. The question therefore remains open as whether the Commission would dare to use the common specifications mechanism. While some experts believe that common specifications are inevitable [P5], others do not believe that the Commission will resort to them [P9]. For [P5], CEN-CENELEC's failure will only become apparent at the last minute, when the Commission will have its back against the wall and it will be too late to draw up common specifications for the deadline. Common specifications will therefore been developed in a hurry.

[P6] admits that nobody knows what these common specifications will look like or who will write them. Although the Commission will be responsible for adopting them by means of implementing acts, it is still unclear who will draft the technical specifications. The Commission could draw up these standards itself, with internal experts, or ask outside parties to do so on its behalf. However, several experts believe that the Commission does not have the necessary expertise to draw up such documents itself [P9, P4]. [P9] explains that most of the Commission staff currently participating in CEN-CENELEC's work today, are for the most part policy experts, and do not have the necessary technical knowledge. The experts are therefore convinced that in the event of common specifications, the Commission will simply ask the same JTC 21 experts to draft the common specifications, but in a different con-

[51] According to our definition of technical standard given in Chapter 4, Section 4.2.1.

[52] For more information about the role of common specifications, see Chapter 7, Section 7.9.4.

text. For [P5], it would be a solution to avoid lobbying, by selecting only trusted experts. On the contrary, [P9] warns that all the usual standardisation procedures, in particular the consensus system, and everything that gives a standard its legitimacy, will be abandoned until standards are published.

8.6 Areas of agreement and disagreement between stakeholders

Many experts are highly critical of the way standardisation work. When asked about their frustrations or what they would change if they had the chance, we got a wide range of answers. When we further asked how the standardisation system could be improved, experts came up with imaginative solutions. However, it should be noted that the experts' experience of standardisation varies greatly according to the stakeholders group they represent, their background and their previous experience in standardisation. In this section, we attempt to summarise this wide range of experiences and the way in which the experts describe the difficulties they encounter and how to overcome them.

8.6.1 Are the processes too complex or do the new experts just need time to understand them?

The problem most often mentioned by participants is the complexity of the processes put in place by international and European organisations for proposing, drafting and adopting standards.

A new vocabulary to learn

The first difficulty is actually registering in a NSB. Some experts, who first took an interest in standardisation through AI and are therefore fairly recent to the standardisation world, tell of their difficulties in understanding the different membership statuses and deciphering the membership contract [P8]. Others found it difficult to join without paying the fees, even though they were among the exceptions for which the fees were supposed to be waived. Some experts advise using an insider, someone from the same NSB, already registered, who can help through the steps and act as a point of contact within the NSB [P2].

But the real difficulties begin at the first meeting. Newcomers receive no assistance and discover a whole world with its own language and customs [P8]. [P10] recounts: *"You find yourself in the meeting room and you don't know nothing about how it works and you hear they talk about numbers, they talk about documents, and you say: 'what's going on?'. This is a very common situation. [...] It is kind of alienating because it is very weird that you are in a room and there are maybe other thirty people you don't know nothing about. And then there is a form of rituality at the beginning of each meeting where you need to introduce yourself and read the Code of Conduct and so on"*. The acronyms mentioned by [P10] are used everywhere, from the names of NSBs and SDOs, to document numbers, to the titles given to the stages in the drafting of a standard. It can thus prove challenging to understand the structure and the role of each groups and each person. Unfortunately, this vocabulary does not have to be learned once, but several times, as even if the general operation of the SDOs is fairly similar from one to another, the vocabulary used changes slightly. *"What is difficult is that there is the vocabulary of ISO, the vocabulary of IEC. There is the vocabulary of JTC1, which is the* *meeting between ISO and IEC, which is again different, an in-between. And then there is the CEN-CENELEC vocabulary, which is again different from all that"**, [P5] says.

Even the simple search for information can prove a gruelling task. Everything is tracked, minutes are taken of every meeting and every document, every proposal, roadmap, framework, is compiled on the online platform provided by CEN-CENELEC. As a result, hundreds of documents are uploaded to the platform with little to no sorting [P4].

A world with its own rules

Once a new expert has started to understand which group they can get involved with and for which tasks, they still have difficulty understanding how to get their ideas across. First of all, the standards development and voting process is highly complex. There are stages to go through before sending in proposals for standards or comments, and deadlines to meet. In addition, these rules may differ from one SDO to another [P9]. The adoption of standards is a case in point. Whereas at international level, one country equals one vote, at CEN-CENELEC, votes are weighted according to the importance of each country. In addition, there are several levels of voting where non-European countries are always invited to participate, but their vote only counts if the result of the vote between the other countries is not clear. The voting system is so complex that CEN-CENELEC has developed a tool to count the votes and calculate them automatically. *"To understand, you need a diploma"**, [P4] says ironically. Some experts, who have been involved with CEN-CENELEC for a long time, admit that they still do not understand all the processes [P9].

A standard way to draft standards

Even when processes are known, there is a certain way to draft standards for them to be accepted [P3]. There are writing customs when it comes to standards which, if not respected, result in the proposal being rejected. Experts have to learn over time, by trials and errors, or by taking inspiration from other standards. *"You imitate a lot. You look at other people's standards and the way they are written, especially those of people with more experience than you"**, [P9] explains. [P9] further recounts that when they first started at ISO, an expert with decades of experience in the field of standardisation sent them hundreds of comments to correct grammar, typos and forms that did not correspond to ISO's writing style. These writing rules are compiled in a document published by ISO, the "ISO House Style", to guide the language, formatting and presentation of ISO documents (ISO, 2024). *"It is just editorial guidelines. It is a misuse of a particular verb, a formatting error, a comma you are not allowed to use here... It is unbelievable"**, [P9] says. But some rules are more implicit. [P3] remembers that when they first started out, they always quoted their sources until another expert told them that they were going to get into trouble: if they quoted too much, other experts would go look into the quotes and they would find something that was not completely clear to discredit the proposal. In the end, standards are supposed to set out requirements, not explain where these requirements come from [P3].

Processes slow negotiations and deter people from coming

Experts report that these heavy processes sometimes get in the way of efficiency. Some standards took years for the proposal to be accepted and the real work to begin. The numerous back and forth, the negotiations, the time it takes to vote on proposals or agree on a few wordings, are sometimes seen as a waste of time. [P8] recalls: *"To launch the trustworthiness standard proposal, it took two years just to write two paragraphs"**. Yet, for hENs specifically, the publication schedule imposed by the EU Commission is very short. The lengthy processes of SDOs therefore seem ill-suited to such urgency. This time spent on processes rather than meaningful discussions is even worse when SDOs try to cooperate, as the parallel development of standards by ISO and CEN requires voting and agreement on proposals from both sides [P9]. For [P8], these heavy processes were suited to in-person meetings but do not work well online. *"We suffer from processes that were imagined and defined in a different world"**, [P8] sights.

But most importantly, these complex processes deter people from investing time in standardisation, because the learning curve is steep, especially at the beginning. People come at a first meeting, then get discouraged and do not come back [P5]. For [P8], if experts miss a few meetings because of other obligations, they can easily get lost and this dissuades them from participating again. [P2] recounts: *"To enter the world of standards, you have to be a bit of a masochist. [...] People talk to you with numbers or acronyms all the time. [...] The effect of speaking in acronyms blocks the possibility for others to understand. So you have major decision-making power because you use the acronym"**. The complexity of the processes and vocabulary sometimes seem designed to prevent people from entering the world of standardisation. It acts as a kind of competence test to prove one's worth. Standards can therefore only be developed with stakeholders who are already familiar with this world. The same applies to the drafting of standards and the writing rules imposed by SDOs, excluding those who do not know the customs from making useful contributions. *"It is like playing a board game: if you do not understand the rules, you cannot play"**, [P11] summarises. [P15] explains that experts have to fully commit and invest a lot of time at the beginning if they want to be able to understand anything. Coming once or twice does not allow you to get used to the jargon. This necessity to invest time to comprehend this world and its codes hinders participation, when there are already few experts actively involved.

A question of experience?

The complexity and slowness of standardisation processes is a divisive issue. While these processes are widely criticised by experts who started standardisation with AI, long-standing industry experts explain that these processes have been written and refined over the years, now reaching a high level of maturity. In their view, all these processes and rules have been put in place to ensure the smooth running of the system and must not be tempered with. When asked about the problems of standardisation, they blame other experts who, in their view, do not respect these processes or the chain of command [P12]. We note here that there is probably a generational conflict between the new experts and the older ones[53].

[53] We are not using the term "new" here as a synonym for "inexperienced", as many of these experts already have several years' experience in standardisation, but to emphasise the fact that they started standardisation with AI, compared with more established experts who have sometimes been in this world for several decades.

8.6.2 *AI standardisation needs to attract more experts... But who?*

Although the work of standardisation experts is extremely important, the job has many drawbacks. In particular, there is a shortage of experts, which makes the work more difficult and stressful for those who stay.

Very few experts are actively involved

Usually, when registering with the NSB, experts undertake to participate in standardisation work at European or international level, although there is no formal obligation or level of involvement required, as well as no consequences for non-participation. Experts can generally enter and leave discussions as they see fit, even if this is not necessarily welcomed by the other more invested stakeholders.

Very few experts are therefore actively involved in standardisation, compared to the hundreds of people registered. Although the data from CEN-CENELEC shows hundreds of registered experts, [P5] estimates that around fifteen people are actually active across all working groups. *"You end up with standards that are actually made by very few people"**, [P8] complains. The other registered experts do not participate in the drafting of standards, but simply carry out monitoring work. They observe the work being done, gather information and report back to their companies [P10]. This enables industry players to keep abreast of forthcoming regulations and stay one step ahead of the competition. From time to time, they attend a meeting and remain in the background without speaking, but most of the time, they simply monitor the emails sent by the SDOs [P5]. [P5] explains: *"Some people are only there to inform their boss. [...] It provides a quick access to documents"**.

But this behaviour is not to everyone's taste. [P5] explains that it makes it difficult to raise the alarm about the lack of experts in standardisation when figures show that hundreds of people signed up to participate. In addition, registered stakeholders advertise their involvement in standardisation on social media and apply for various projects – in research or industry – highlighting their standardisation expertise, when in fact they have never actively participated in working groups [P5].

The large number of experts with little involvement also makes it difficult to move discussions forward. As [P13] explains, the subject of AI attracts many people who come and go between discussions. In particular, some experts arrive well after the work of a working group has begun, do not necessarily look at the history and reopen debates that were dealt with long before their arrival. As a result, the group is forced to reach a new consensus on a subject that was supposed to be closed, wasting time and efficiency.

Involved experts have a heavy workload

To make up for this lack of experts, some of them try to get involved in all the working groups, forming a *"hard core"* of two or three people who are present everywhere [P12]. These experts explain that they are very involved because the working groups are short of people and they feel obliged to fill the gaps [P5]. The involvement of experts therefore varies greatly, from people who only read emails or go to meetings to listen, to people who focus on a specific working group and devote all their time to it, to people who are part of all the groups at once. Experts' workloads thus depend on their level of involvement, as well as their working groups. Some groups are more active than others, with editors receiving dozens of comments and several contributions a week. On the other hand, other groups, particularly those working on standards that will not be harmonised, are struggling to attract participants [P10]. Different experts therefore report various working experiences.

For people who only focus on one topic, it can be manageable to keep it as a side activity. [P9] told us they devoted 300 hours a year to their working group, so about one day a week. But even a few hours a week can be a lot for experts who are not supported by a structure or company and who have to invest time in addition to their usual workload. *"It is interesting, but it is not supposed to be my job, so I am working overtime"**, [P5] confesses. In periods of heavy workload, a working group can meet weekly for around two hours, or even four hours for some. Depending on the activity of the groups and the number of groups in which the experts participate, the workload can vary from two hours a week to a full-time job [P12, P13]. Indeed, experts often have to work outside these meetings to keep the project moving forward. [P5] acknowledges that with just one or two hours a week, the work does not go very far, and that the experts need to invest more to make it worthwhile. Thus, some experts, who are supposed to work only part-time in standardisation, report that they actually spend much more time on this than their official quota [P10, P15]. [P3] also notes that the workload has increased since the standardisation request. [P12] does not welcome this increase in the number of meeting hours, which for them, does not mean that the group is necessarily more productive. Indeed, meetings can last for hours, sometimes over details, just so that a proposal can be rejected and everything needs to start over. For the experts that strives to be everywhere, all at once, meetings can represent up to 11h a week between SC 42 and JTC 21 [P5]. With meetings sometimes taking place at the same time, some experts admit they sometimes had two computers connected to two online meetings at the same time so that they could follow everything [P5]. Furthermore, as the ISO secretariat is located in the US, these meetings can take place very late at night for Europeans trying to keep up with ISO's work [P5, P15].

This investment in time, work and sometimes hours of sleep, can further deter people from coming. As [P11] explains: *"Recruiting people for whom this is not their main objective or main job is not easy, especially when you explain to them that they will have to get involved, but not lightly. It is a three-year cycle, at least, and you have to contribute. You cannot just be there one time and gone the next"**. Standardisation thus seems to be stuck in an eternal cycle, where fewer experts means more workload, but where more workload also means fewer experts.

Academic and SME experts have a hard time finding funding

Another reason keeping people away from standardisation is money. Indeed, standardisation work is voluntary work as people are not directly paid by standardisation organisations to work on standards. Nevertheless, this does not mean that experts are not paid at all. When they work in a company, monitoring discussions and advancement of standards, or participating in standards drafting, is often part of their job or mission tasks. Some organisations may also receive funding to assign people to standardisation. This is the case, for instance, of European consumer associations and trade unions [P6]. Researchers from public universities or research institutes, on the other hand, do not receive additional funding to participate in standardisation initiatives, which is often cited by academics as a barrier to entry [P5, P16].

First, academic researchers have to convince their university that it is really worth investing time in standardisation, even though there is no product to sell, unlike companies, which have a clearer interest in participating. Second, the travel expenses to go to plenary meeting can be quite expensive as they take place all around the globe. Experts from academia often have to call on special grants to finance their travel to be there in person. The best known of these grants are distributed by StandICT[54]. But according to some experts, these grants are neither sufficient nor well distributed [P5]. Grants are awarded to a few dozen people for the entire standardisation of digital technologies, which in the end does not represent much money dedicated to AI. But most importantly, this income is highly unstable and some well-established experts are often turned down. Grants are awarded by independent experts who do not necessarily have the expertise or knowledge of the inner dynamics of standardisation committees. For academic experts, it is therefore a heavy mental burden to have to apply every time, with the risk of being rejected. On the contrary, [P2] welcomes this extra money: *"It is not much, but it supports our activity"**. To compensate, many experts are multi-affiliated, seeking sources of income wherever they can. This involves teaching courses, working for a consulting firm or a private organisation, and so on. For [P8], this is understandable, as it is not in anyone's interest to register on their own to take part in standardisation efforts. Experts need to be sponsored by other structures to cope as best as they can.

This lack of financial resources creates a gap between large organisations, which can afford to pay for travel expenses and staff working hours, and smaller companies that sometimes have difficulty finding people with time to invest [P13]. But this gap is even more important between these companies and academic researchers, who can hardly use their university money to travel around the world for meetings [P5]. Some industry experts even admitted to us that they did not understand how academic experts managed to still get involved [P3]. For [P8], the lack of experts in standardisation today is therefore due to a combination of the time investment needed, and the difficulty of finding funding to cover the experts' activities: *"It is very difficult to find actors who have time to understand our codes, understand what we do, how we do it. Who has the time for all that, and who can be paid"**.

Academic researchers struggle to get recognition for their work

But apart from money, there are other drawbacks to being an academic researcher who wants to work in standardisation. There are constraints in the academic world, on teaching and producing academic articles [P8]. Although the way in which university systems operate depends on the country, in Europe academics are often asked to devote the majority of their work to research. In particular, they are encouraged to publish in high-quality journals and conferences. A researcher's reputation is therefore necessarily linked to the number of publications, the quality of their publishers and the number of citations their articles receive. However, all of this activity is necessarily reduced when academic experts spend time on standardisation initiatives. Some experts told us that they are putting their academic career at risk by investing so much time in standardisation [P5].

This notably comes from the absence of recognition of individual work as the involvement of the experts on standards is anonymous. Unlike academic articles for which they receive visibility, standards have no return on investment for academic

[54] To learn more about these grants, see Chapter 4, Section 4.6.5.

experts. It is possible, however, to obtain a certificate of recognition of work, even if the expert's name does not appear on the standard, to prove that the expert did participate [P5]. Nonetheless, the criteria for what it means to "participate" are hard to define.

Taking everyone in, regardless of competences

The number of experts actively working on AI standardisation is so low that NSBs are trying to recruit as many new people as possible. This lack of experts also leads those who are active in standardisation work to advertise as much as they can around them, in order to attract more people. [P9] acknowledges that they cannot afford to turn anyone away. *"At this point, I will take anyone who wants to come in. Today, we do not have enough experts. I am not going to start getting picky. There are so many standards in progress that there are entire areas where we are totally absent"**, explains [P4]. The lack of selection on entry to NSBs can also be explained by their business model, in which anyone who pays the price, i.e. the membership fees, can participate in standards setting.

But this lax approach to selecting new entrants raises the question of the real competences of standards experts. *"I don't know who is competent or not"**, admits [P4]. As a result, people with no previous experience of standardisation, or of AI on a technical standpoint, are nevertheless sometimes quickly promoted to editor or convenor positions if they show an interest in a given subject [P16]. *"If people pay, it means they are motivated, have skills and something to contribute. There is zero selection. After that, it is a question of collective intelligence"**. [P9] reckons that it is still quite rare to have people who are totally unfit, but that it can happen from time to time. *"I have been told by people who had joined a NLP [55] working group, 'I am coming because then I will know what NLP is'. It is a shame that NLP regulatory standards will be written by someone who came to find out what NLP is"**, recounts [P5]. [P3] thus question the use of the work "expert", claiming that people involved in standardisation are not really experts in anything[56]. Some of the experts actively involved in AI standardisation recognise this themselves. *"I am technologically obsolete"**, says [P11], a "governance" expert, involved in various standardisation activities but who is not an AI specialist.

However, beyond technical expertise, knowledge of standardisation processes is essential for experts who want to get involved. [P5] admits that even if some people do not have the necessary background to take part in technical discussions, they can still contribute to the strategic aspects by setting up cooperative ventures or organising the work. Nevertheless, problems arise when non-technical experts take over technical subjects. The content of standards then runs the risk of becoming more political than technical[57].

Computer science experts do not feel legitimate

On the contrary, it is often computer science experts that do not feel legitimate to work on standardisation. Some start the interview by insisting on the fact that their background is not in standardisation [P7]. Others say that they do not feel competent in matters of standardisation, even when they have been the main contributor to a standard [P3], or say that they should not have led a standard because of their lack of competence [P16]. Finally, some experts repeated several times during the interview that they had only been involved in standardisation for a few

[55] Natural Language Processing

[56] As a reminder, we have chosen to use the term "standardisation experts" because it how the SDOs refer to the individuals working on standards (ISO, n.d.i). For more information, see Chapter 4, Section 4.5.2.

[57] See Section 8.6.4.

years [P15]. Individual legitimacy therefore seems difficult to build in the world of standardisation.

This phenomenon can be explained as, contrary to what their name entails, technical standards are rarely very technical. This is specifically true for AI where standardisation work is just beginning and first works are necessary more high-level. But writing high-level requirements is something that technical experts, such as researchers in computer science or engineers, are not used to do, and that they do not often like to do [P8]. *"Writing requirements like that, this is not our job"**, says [P8]. People are impressed, they see standards as a particular type of document that they are not qualified to write. Even when technical experts manage to take the first step of registering to participate in standardisation efforts, they are greeted by discussions in which they do not feel confident, which are more strategic than technical. *"I felt like I was walking into the Assembly's legislative process"**, jokes [P3]. Computer science experts in particular can get frustrated, as *"economic, strategic and business incentives"**generally take precedence over technical discussions [P15].

The reticence of technical experts to work on standards may also be due to the global standardisation system based on national structures where experts represent their country of origin. [P8] believes that this sovereign mission is what drives certain expert profiles towards standardisation, but it is also what puts off more technical profiles. In these organisations, knowledge of the system itself is sometimes more valuable than technical knowledge. This is even truer for European organisations, where standards have an intricate relationship with the law. [P14] admits that for their work at CEN-CENELEC, they use their legal knowledge of the AI Act and the New Legislative Framework more often than their technical knowledge of AI.

Personal difficulties depend on the expert's stakeholders group

It is undeniable that there is a great difference between the experience of academic, civil society and SME experts in standardisation and that of larger companies experts. Whereas academic experts often find it difficult to reconcile their career goals with their standardisation activities, working overtime and struggling to find funding, experts employed by a company are paid to follow standardisation initiatives, even if this is not always a full-time assignment, which facilitates their participation. Industry experts therefore have more time to contribute or take leadership roles. Academic researchers involved in standardisation often find themselves more isolated than industry experts. Even when they have benefited from an inside contact that has brought them in, they have no formal links with anyone and have no specific interests outside their own. They often have to discover this world and its rules for themselves and struggle to fit in. On the other hand, experts from big companies often already have a colleague or manager working with them and, depending on the size of the company, may even have a team of several people to work with. They subsequently feel less isolated.

8.6.3 *International influences: saving time or threatening sovereignty?*

ISO and CEN are accustomed to cooperating under the Vienna agreements. But this cooperation, and in particular the adoption of international standards by Europe, may raise questions of sovereignty.

ISO and CEN: a continuity of the same world

ISO/IEC SC 42 and CEN-CENELEC JTC 21 do not operate in the same geographical area. SC 42 secretariat is American and their working groups welcome members from all over the world. On the contrary, JTC 21 is limited to European countries and some affiliates[58]. However, many of the experts we interviewed pointed out that the same people actually work on both committees. Indeed, the experts who are truly invested, and do not only monitor the activities of the committees, usually participate to both SC 42 and JTC 21 [P2]. This include the few "core" experts who are omnipresent and who take part in almost all working groups [P3]. [P9] estimates that about 80% of JTC 21 experts are also in SC 42. It seems, however, that the international level gather more experts, thus completing the overall picture alongside Europeans. But European experts are only a subset of international experts. The only exception is consumer organisations and various European associations, which have the right to participate directly in CEN-CENELEC but, as they are not part of any NSB, have no say in international standardisation [P11].

Nevertheless, the connection between SC 42 and JTC 21 does not stop there. First, the two committees use the same IT tools, such as the same login for experts, which is shared between the two online platforms. *"You see, it is basically the same world"**, concludes [P8]. But above all, they often work together, thanks to the Vienna agreements, which enable them to adopt each other's standards and launch joint standardisation initiatives. When two groups from the two committees appear to be dealing with related issues, they officially appoint a liaison officer to represent the group at the meeting of the other committee. It is then possible to share documents, talk about future standards initiatives, ask for comments, and so on [P9].

ISO experts push for international standards to be adopted at European level

As many experts participate in both ISO and CEN committees, they prefer not to work twice on similar subjects. Since ISO is ahead of the game in its standardisation work[59], many experts are advocating the adoption of ISO standards at European level rather than the development of new standards. According to [P8], these experts, who are also involved in international SDOs, come to CEN-CENELEC saying that ISO is already working on the subject and try to dissuade the European experts from launching a group on the same topic. For [P2], this avoids *"reinventing the wheel every time"**. It is indeed less costly for the companies involved in drafting the standard to have a single standard to develop [P15]. It is also easier for companies to comply with a single standard than with a multitude of different standards. So, at the same time, these experts are pressing for international standards to be adopted without modification by JTC 21 [P6]. For [P5], the experts working at ISO have an interest in turning JTC 21 into an *"empty shell"**that simply enacts what has been decided at ISO.

But these efforts are coming up against resistance from European groups. *"ISO is trying to bypass us on European standards, but they are not succeeding"**, says [P4]. These victories are, however, not perceived as such by everyone. [P6] points out that many ISO standards have nevertheless been adopted, or are in the process of being adopted, at European level, only they are not yet on the way to becoming hENs. But this possibility is not entirely excluded.

[58] For more information on who can participate in CEN-CENELEC, see Chapter 4, Section 4.5.2.

[59] For more information on the work of JTC 21 and SC 42, see Chapter 5, Section 5.3. For a discussion over the "gold race" to AI standards, see Chapter 7, Section 7.6.

Adopting ISO standards poses issues of sovereignty

However, the adoption of international standards by European committees raises issues of sovereignty. Europe wants to retain control over its standards, particularly when they are to be used to support legislation, as is the case for hENs under the AI Act. [P8] discusses this possibility: *"As a result, the European standard that offers a presumption of conformity is in fact an ISO standard that was developed in collaboration with the Russians and the Americans, and even without taking European regulations into account, either because it has a broader scope, or because that was not its mission anyway. The standard may even have been developed before the AI Act"**. In fact, ISO standards were not developed with a view to implementing the AI Act, but for economic reasons. The lack of AI standards until 2018 had simply created a demand and therefore a market. *"They don't care about the IA Act in international organisations"**, adds [P6]. As such, many ISO standards are not fit for the AI Act [P15]. For [P2], this reluctance to adopt ISO standards can be explained by the fact that in Europe, there is a focus on fundamental human rights and European values, which are not shared at international level. This can be seen, for example, in the emphasis placed on fundamental rights in the AI Act and recent initiatives to include ethical and fundamental rights issues in European standards[60].

Among the ISO standards that have been adopted at European level is ISO/IEC 22898, the *"bible of AI terminology"**[P5], which compiles all the terms relating to AI. For some experts, this prevents European players from having a say on the definitions of terms that will ultimately influence European standards. [P5] explains that it is impossible, for example, to deviate from the definition of "transparency" imposed by this ISO standard, even if European groups decide to tackle this issue. A standard on transparency will then have to be aligned with the definition in ISO/IEC 22898 when defining requirements. This poses problems when the definition does not correspond to European political interests. Indeed, today, the definition of transparency for AI refers exclusively to the transparency of outputs and not to the transparency of the system itself. According to [P5], this definition suits companies who do not want to give access to their system. Introducing the notion of transparency of the system into the standards could make it possible to challenge companies on this notion when reference is made to "transparency" in legal texts.

For [P9], the roots this problem lie in the Vienna agreements: *"I am caricaturing here, but [the Vienna agreements] say that Europe voluntarily relinquishes its sovereignty to ISO as long as ISO considers that they are interested"**. But for [P11], who is also engaged in other standardisation activities at ISO outside of AI, the adoption of international standards is a necessary concession, as ESOs have no time to do otherwise given the deadlines imposed by the Commission with the AI Act. JTC 21 therefore needs to keep moving forward and stop looking in the rear-view mirror. [P11] recognises that there is a lack of confidence in international standards, but, to them, it is not justified.

Opinions are therefore divided on the benefits of ISO standards for European standards. While many of the experts who initially worked on these ISO standards are in favour of their adoption in Europe, including as hENs, new experts usually want to renegotiate from scratch, which could delay standards but would give them a more sovereign approach.

[60] For more information, see Chapter 5, Section 5.2. To have the opinions of our experts on the feasibility of such initiatives, see Section 8.5.2 of this chapter.

Capturing the interest of European companies

For some experts, this mistrust of international standards can be explained by the fact that many of the companies involved in standardisation are American or Chinese BigTech companies. On the contrary, European companies are virtually absent from the discussions. While this is already true at European level, there is a balance brought about by the participation of consumer associations who are absent from international groups. For [P11], sovereignty issues could be addressed through the participation of more European companies in standardisation, whether at international or even European level. However, [P9] notes that this strongly depends on the country. In Germany for instance, there is a strong culture of standards and companies are well aware of their importance. On the contrary, this is not the case in France. According to [P11], it is impossible to compete with American or Chinese companies, because European companies are not as powerful. On the contrary, [P9] believes that in standardisation, all the players, even the smallest, can have a seat at the table. It is simply a question of investment.

And indeed, many big European companies could get involved if they wanted to, because the human cost is not much compared with their number of employees or their investment in other areas. But they do not get involved because they cannot see the economic benefits. For [P4] it is because there are no visible short-term interests and they fail to reason in the long term. *"It is a cruel lack of strategic vision. [These companies] are incapable of seeing the return on investment of their activity beyond a quarter or a year. When we tell them 'No, but you have to write the standard now, so that in 3 or 4 years' time, you will dominate the market because your standard will be the rule', they say 'that's too far away, too uncertain, what counts is now"**, explains [P9]. There is a paradox in that some companies are involved in major partnerships and collaborative projects that are quite similar to standardisation work, but they are not directly involved in standardisation. *"We have large companies who find it hard to understand, who say 'we will see, we will manage with the standard'. You are joking. You invest tens of millions in robustness and you are not even involved in drawing up the standards that will tell you what robustness is. Maybe everything you are doing will be thrown in the bin"**, laments [P4].

It is difficult to quantify the return on investment, but it does exist [P10]. *"When you write a report or do your annual review, you find it hard to say 'I helped save this much' or 'I have produced this much', because it is not quantifiable. You have certainly saved your company money because you have influenced a standard that could have been negative for them or that could have had a financial impact on the way the company operates. But you cannot quantify it like that"**, [P11] explains. It is therefore necessary to make these European companies understand what this return on investment can be. *"You do not work in standardisation purely out of altruism. [...] There must be an interest, and very often a financial interest"**, [P11] adds. For [P4], it is therefore urgent to encourage these companies to participate in standardisation, because on the contrary, GAFAM and BATX have perfectly understood the long-term benefits of participating in standardisation and are making their voices heard.

8.6.4 The presence of industry: an opportunity for practical implementation or a risk of control?

The issue of industry participation in standardisation is already well documented in the literature[61]. Here, we have given the floor to non-industry experts to explain their experience of working with companies, and to industry experts to defend the benefits of their participation.

A large representation of the industry

All our experts, without exception, mention the strong presence of industry in standardisation, even if this is not perceived in the same way by everyone. The data we collected and analysed confirms that industrial players are over-represented among the experts. *"Yes, but that is part of standardisation. That is what standardisation is all about. It is about big companies getting together and trying to reach a consensus on the same thing"**, explains [P6]. Standardisation committees are therefore contested arenas, as they are highly prone to lobbying [P7]. *"The Commission knows it, everyone knows it. We just deal with it"**, admits [P4].

This can be explained as the industry, specifically bigger companies, have the man power and money to allow for meaningful contribution. They can put people full time on a topic and pay for their travel expenses. They do so because they have financial and strategical advantages to participate. While some consider that this lobbying is not beneficial to standardisation, industry actors believe that they bring expertise that no-one else has. *"What is criticised is the predominance of industry. But the industry is the driving force [behind standardisation], both through innovation and through the resources it is able to deploy"**, [P11] explains. But the human resources deployed by industrial actors, go far beyond a simple help in standards development. According to [P5], they control many topics of interests, decide when – or if – initiatives should be launched and what the content of the standards should be.

Transnational corporations make use of the national system

Large companies are implementing strategies to influence the development of standards. Several experts explained to us that the strategy of these companies for controlling the system is to make use of the national structure of standardisation. Because they are transnational companies and have staff in different countries around the world, they can claim to enlist experts in several NSBs. Once their experts are established within the different NSBs, they exert a form of influence in each country [P2]. This influence is even greater if an expert from these companies manages to get themselves elected head of delegation, in order to control a country's position at plenary meetings. *"It is easy, when you are making tens of billions in profits every quarter, to find fifty people to infiltrate the whole system of European standardisation. [...] They have a huge presence, and they are the ones determining the position of countries"**, explains [P4]. They coordinate between experts from different countries, but from the same company, to push forward the proposals that suit them and oppose those that they do not want. [P5] recounts: *"I presented [my proposal] and, within 10 seconds, seven hands went up. These seven people said 'I object, I object, I object...'. They were seven people from different countries, who had nothing to do with each other, except that they worked for the same company, for the same team leader"**. This experience is not an isolated event, as several experts report having witness the same

[61] We discuss it in Chapter 7, Section 7.7.

type of behavior [P4, P6].

The consensus mechanism that underpins all standardisation decisions gives these companies an advantage, because if one of the experts finds reasons to oppose, the whole process can be blocked. A proposal is therefore only accepted if these industrial players are willing to accept it. *"We say it is the consensus, but really it is the consensus of people with interests"**, summarises [P4]. Exposing these experts is also made difficult by the fact that experts have no obligation to display their employer during plenary meetings. When experts are confronted with such a situation, they cannot prove that it was a corporate strategy to boycott a decision [P5]. The standardisation system is therefore very favourable to these large companies, which do not hesitate to use their power to defend their interests. *"Standardisation is the preserve of some big companies which have perfectly understood how it works"**, says [P4]. According to [P8], some experts have even left the world of standardisation because they were tired of this unwinnable battle against corporations.

Some companies have no interest in publishing standards

Industry experts hold the keys to standardisation because they have the numbers and the strategic positions in each national organisation. They often try to block standards, because it is in their interests that they are not published [P2]. Some experts point out that companies generally start by saying that such a standard is not necessary [P5]. When this works, it avoids any negotiation. But when they cannot completely block a standard, companies usually manage to keep the discussions going for longer, in order to gain time. *"There are other strategies that involve nitpicking every word, dragging things out and having the standard arrive three years later"**, [P4] explains. [P15] describes the strategy of these actors who prefer to block the standard: *"All I have to do is [spend] three hours on a paragraph. Nobody has the right to tell me to shut up or anything like that because I can always come back. 'No, no, I don't like that sentence'. Instead of moving on, I block it out. At a certain point, I scare off the experts. And if the experts get fed up and leave, [the standard] will never be finished"**.

Some experts, particularly those with a background in computer science, get easily bored in these interminable discussions. [P5] reports having spent hours in working group meetings just to look at an Excel sheet. For [P3], these companies are dishonest, because they do not accept any propositions. There is no negotiating ground.

Some of the experts we interviewed believe that there is a fundamental paradox in standardisation, because the main contributors have an interest in not publishing these standards. As [P9] points out, unlike other fields, standardisation of AI is not a necessity. Indeed, the Telecommunication industry needs standards because phones which cannot call other phones are useless. On the contrary, today, while it is sometimes necessary to be able to integrate AI into downstream applications, most of the time AI systems can be used on their own. This reduces the incentive for companies to collaborate in developing standards. [P3] explains that this is what happens in certain working groups, which are dormant because they are mainly made up of industry experts who have no interest in promoting a standard. Standardisation then enters a game of appearances in which industry experts are forced to show that they are contributing, but without doing too much at the same time.

However, as [P5] points out, the existence of hENs primarily benefits SMEs, as large companies have the financial means to pay for alternative compliance frameworks. If the attempt not to publish standards succeeds, it is therefore small businesses that will pay the price.

Some companies push for empty standards

Often the main concern of the industry is to ensure that the standard does not contain too many strict requirements that would hinder the companies' economic interests. According to experts from academia, consumer organisations and SMEs, large companies have a habit of blocking technical contributions to ensure that standards remains as vague as possible. [P5] explains that the general strategy is to prevent other experts from making significant contributions. For example, to avoid introducing metrics into standards, industry experts claim that the field is not mature enough [P5]. They also advocate for more guidance and fewer strict requirements. To achieve their ends, they would for instance submit comments when drafting the standard, which try to get rid of important keywords [P9]. Certain words which appear in the AI Act can then be left out of the standards. As a result, the standards are not fully operational [P4]. [P5] cites the case of ISO 5259 on data representativeness, where the definition goes around in circles, defining representativeness as the number of representative elements in a dataset. Similarly, "performance" is defined by ISO in standard 22898, simply as a "measurable result". As a result, these standards are useless to complement legal texts.

Another widely cited example is ISO/IEC 42001 (2023b) on management system for AI, inspired by other similar standards, such as ISO 9001 (2015b) on quality management systems. Like 9001, 42001 aims to standardise a company's processes, not its products. Therefore, it does not contain too many technical details. For [P1], this is a necessity, in order to be as independent as possible of technologies or practices which could become obsolete. [P1] explains: *"[In 42001] we say: 'the organisation must define its performance requirements, document them, verify them, demonstrate that it has indeed achieved the requirements it has set itself' [...]. But we are not going to set the thresholds"*[*]. The certification audit subsequently verifies whether the company has actually achieved the objective it set itself [P1]. Although management system standards are a great success in companies all over the world, some of the experts we interviewed are very critical of the approach of this type of standard, which they feel is not strict enough [P6]. They feel that these standards leave too much to the discretion of the company, making certification pointless. For [P5], there is no doubt that the 42001 standard does not say much because the main drafter is the industry itself. These large companies will then comply with this standard which they helped to draft, in order to obtain a certification mark that they can use for marketing purposes.

[P5] further warns that standards on management systems are not designed for smaller companies and could kill businesses if they were imposed on everyone, as if 42001 became a hEN. According to [P13], an expert from an SME, when it comes to standardisation, it is necessary to differentiate between the interests of large companies and those of smaller companies, which generally prefer to receive specific instructions on what to do. *"There is industry and industry"*, says [P13]. SMEs are generally more specialised and therefore have more specific needs. For example, developers and deployers of AI systems will not necessarily have the same incentives in standardisation. However, SME experts report that BigTech companies, who operate at an horizontal level, across sectors, do not listen to companies who are operating at vertical level [P9]. For [P13], it is important for the SME's voices to be heard, as today there is *"too many of one voice"*, that of larger companies.

On the contrary, some experts, even outside the industry, told us that if standards were to set requirements that were too precise but ill-adapted, this could kill the industry [P16]. It is therefore difficult to find the right balance between requirements that are too precise and requirements that are too high-level, particularly when the standards are horizontal, such as the AI Trustworthiness framework, and it is therefore not possible to rely on contextual elements.

To summarise, standards are not neutral. [P4] explains: *"When you don't have this critical view of how standards are developed, you take them as they are, without questioning where they come from. But there has been a whole process and some people have managed to ensure that the standard does not hold them back. Standards are a tool of strategic and economic warfare. People don't understand that. Given the cost, standards are not developed by philanthropists. They are made by people with economic interests. And some people have an interest in the standard being empty. [...] Everything they have written is in line with their product policy. There is a whole business model behind it"**. But this emptiness in standards is even more problematic when standards are intended to become hENs. [P4] explains that this a corporate strategy to influence regulation: *"Companies do not really have a say in the drafting of legislation, but they do in the drafting of harmonised standards. By putting blurry, empty, hollow, vague things, and not things that are too precise, it is possible to screw up harmonised standards"**.

Different entities have different strategy to influence standardisation

To gain the upper hand on standards, many companies are developing a standardisation strategy. Large companies therefore have a "chief standardisation officer" or a similar role, specifically dedicated to organising the groups of experts working for the company in order to define clear objectives and voting strategies. For [P11], it is all about placing one's pieces on the chessboard. For companies, it is important to know the stakes for each standard and decide whether to get involved or not, and what to contribute. For instance, experts report that they were asked to focus all the efforts on the standards supporting the AI Act [P3]. This is why so many people in standardisation are just dormant and not participating actively in standardisation efforts: they monitor the group for the company and alert their hierarchy when a topic of interest is being discussed.

But strategies vary from company to company. Some say they only have governance experts scattered all over the different working groups. [P12] explains how it works in their company: *"In our profiles, we are 100% dedicated to standardisation and we can cover several subjects, or fields, such as contributing in cybersecurity as well as AI. This means that we are not extremely specialised or in-depth specialists in technical subjects. But we are technical enough, obviously, to understand the subject and what is at stake. More importantly, we are going to take the know-how that interests us within the company, and take positions and make contributions based on that know-how"**.

Other companies, on the other hand, concentrate on a few working groups and send in technical experts who can make contributions in line with the company's objectives [P3]. They are specialist engineers, computer scientists or machine learning researchers, already employed by the company, but who have never set foot in standardisation. They are then asked to take the lead on a specific standard that is deemed important for the company.

Civil society groups are also using a precision shooting strategy to organise their standardisation work. They have experts who lead the team and help coordinate it, while others take an active part in the discussions and make contributions. These groups prioritise notably standards on risk management, trustworthiness, fundamental rights impact assessments and conformity assessments [P7]. *"One of our strategy is 'contribute, contribute, contribute'. [...] You want to be on the table to be heard. [...] I think that is a big part of it really, just trying to make sure that your agenda is in the mix. You cannot determine the end outcome quite so much. [...] But the more you contribute, the more influence you have"*, explains [P7].

Another strategy often mentioned by the experts during the interviews, is to hire consultants [P5, P8, P11, P16]. According to [P5], this strategy is mainly being implemented by Chinese companies to counter the standardisation strategies of American companies. Even when asked to state their employer, these experts may simply display their consulting firm, thereby concealing the client who is actually paying them and thus the company whose interests they represent. *"It is even worse because it is not transparent. The guy comes to a meeting, he is there, he works as a consultant, but you don't know who he works for"**, explains [P11]. This strategy is more expensive for companies, but it is more discreet and gives them rapid access to technical experts. *"The Americans do it too. Except that the Chinese do it even more, because they pay double, or triple, or quadruple. [...] They have unlimited budget"**, adds [P11]. Experts recount that Chinese companies try to hire people from American companies in order to recruit more experts in standardisation [P5, P11]. Ultimately, the more companies invest in standardisation, the more powerful they become in this field.

Balancing companies and national interests

One of the biggest questions in standardisation is the extent to which experts represent the interests of their country as opposed to those of their home institution. For instance, some experts work in American or Chinese companies but in a European country. Therefore they represent the interest of their company, of their country and of Europe. *"You have to be a bit schizophrenic"**, admits [P11]. *"You have to remember who is paying you, but at the same time you have a dual allegiance"**. Many experts find it difficult to decipher the reasons behind the actions of other members. *"Who do you actually represent? To what extent is [Expert name] the voice of [Company name]? To what extent the questions he asked us – which were difficult for us – to what extent that was not just him, as an expert in standardisation processes, who was asking them?"**, wonders [P8].

For [P8], what is expressed is above all the employer's interest. [P4] adds that experts tend to present themselves as representatives of their country, but that this is in fact a disguise for expressing the positions of their companies. Some of the industry experts we interviewed do not hide it: *"When you are doing standardisation, you are doing lobbying"**, admits [P11]. Even outside the industry, experts reckon that everyone lobbies to defend their interests, including the civil society: *"if you are a lawyer and you are protecting workers, you will say that workers are not protected enough"*, says [P16].

On the contrary, some experts are very attached to their neutrality, and explain that when they express a point of view, they do so as experts in the field, and not as representatives of a company [P9]. *"There is no such thing as the [Company name] philosophy. In reality, it is a problem of small people who have a very important role. Everyone is trying to get a bit of visibility, to contribute something positive. So I strongly believe that we need to involve all these groups and give them space"**, says [P2]. In reality, experts explain that it is a question of balancing interests. For [P11], it is ultimately about intellectual honestly, although they recognise that it is difficult to make decisions when their company's instructions and their personal convictions are contradictory.

A system designed for industry experts

Despite the will of European institutions and the SDOs themselves to diversify the profiles found in standardisation, and aim to represent the various interests of society as a whole, these groups are still dominated mostly by big industrial players. There is therefore a gap between what is stated in official communications about AI standardisation and what actually happens behind closed doors [P15].

The *"plurality of voices"* [P13] which are supposed to coexist in standardisation and give meaning to the consensus mechanism, seem to be drowned out by the voice of large companies which can afford to participate actively and through different countries. The entire standardisation system seems to have been designed by and for these stakeholders, with specific processes, customs and vocabulary inspired by the management systems of industrial groups. It is a world that new industry experts also have to discover, but one that is much harder for experts from academia or civil society to understand. Standardisation therefore is a sphere where everyone is theoretically welcomed, but where the efforts to get in and stay invested are different for the various categories of stakeholders.

The presence of the industry in standardisation is both its strength, as it is closer to the concerns of stakeholders, and its weakness, as it is more prone to lobbying from big companies. All experts outside of BigTech companies mention the lobbying from the industry as a main problem of standardisation. On the other hand, BigTech experts acknowledge that there is a distrust of large companies, but maintain that their perspective is necessary for standardisation [P11, P12] – an argument supported by other experts, even outside these groups [P2].

The content of standards is another point of divergence, but it does not necessarily follow the same pattern of separation between the industry and all the others. Indeed, while many industry experts support standards that do not necessarily prescribe too many technical measures, experts from other parties can also agree to this. Some experts would prefer standards to focus more on processes and governance, as they feel that technology is not ready for stricter standardisation. Others think that standards that are too entangled with normative issues should not be making specific decisions either. We therefore heard similar rhetoric from industry experts, calling for standards on risk management that could allow companies to decide their own thresholds, and from civil society stakeholders who explained that the only way to guarantee the protection of fundamental rights through standardisation was to demand greater transparency. In both cases, standards remain at a very high level, not defining technical requirements, but requiring companies to make their own decisions and document these decisions.

On the contrary, we have heard opposite arguments from computer science experts, mostly from the academia. They claim that standards should contain as many technical elements as possible in order to avoid loopholes that companies could use to avoid overly restrictive regulation, as is the case with the definition of transparency.

8.6.5 Can we improve the current system?

When asked to consider improvements to the current system, the experts did not propose changes to the processes themselves, but rather ways of reducing their potential negative effects.

Introduce a mentoring or training system

To counter heavy processes, many experts suggested means to ensure that all members are familiar with processes or could bring themselves up to speed quickly when they arrived in a NSB. For [P10], having people knowledgeable about how standardisation works could help speed up processes. This would prevent meetings being disrupted by people asking questions about processes [P10], or people making mistakes when leading work on standards [P5].

A simple way of achieving this would be to organise some form of training when people arrive at the NSBs. In addition to speeding up processes, it would make it easier for people to participate and enjoy standardisation work. *"I think this is a responsibility of national bodies"*, says [P10]. Although some NSBs provide a few documents to familiarise experts with standardisation, such as a list of acronyms and their meanings, they could also organise voluntary training sessions for experts who would feel the need. However, [P10] acknowledges that for smaller NSBs with only a few experts, this might be easier than for France or Germany where experts come in and out everyday.

Additionally to initial training, NSBs could put in place a mentoring system, whereby experienced players could help newcomers to learn the rules of standardisation. This already happens informally, with some experts saying that, when they arrived, someone took them under their wing [P11]. Training and mentoring are solutions that have already been implemented in many companies. However, they require a lot of investment that only big companies can afford [P2]. There is therefore an additional gap in access to standardisation work between member from big companies and the rest of the stakeholders.

Increase investments and recruit staff

Many of the problems encountered by the SDOs, and by the ESOs in particular, are due to a lack of money and a lack of manpower [P11]. *"When there are a lot of [experts], it is difficult for one person to block [the process]"**, recognises [P15]. Greater investment at European level could enable more people to take part, particularly academic researchers, in order to diversify the pool of experts, as requested by the European Commission. Indeed, the money offered by StandICT is praised by many experts, but it cannot be an expert's sole source of income. These investments could also benefit European NGOs. To date, only a few organisations are authorised to participate in CEN-CENELEC. Welcoming more organisations and providing them with the funding they need to play an active role in standardisation would help to level the playing field against the power of the industry. As well as increasing funding, awareness-raising campaigns are also needed to recruit academics [P15]. [P15] also mentions that initiatives could be put in place to promote the company whose experts are actively involved, to encourage more people to come.

In addition, some experts propose that CEN-CENELEC should employ full-time staff to help with projects, not necessarily to take part in discussions on content, but to help the convenors and editors coordinate standardisation work [P8]. They could, for instance, help them organise group meetings, take minutes, and draft standards under the supervision of the experts. [P11] suggested that the European Commission could appoint project leaders to monitor discussions within the working groups, and speed up the development of standard, particularly with a view to the AI Act.

Impose more transparency during meetings

Another idea put forward by our experts is to have greater transparency within the standardisation system. There are already a large number of ways of tracing each decision, but for some, examining the minutes and comparing decisions is more like an investigation to put the pieces together than clear and direct information. To complete this system, JTC 21 experts have proposed during a meeting to record future meetings, at least the plenary sessions. However, this idea was rejected by consensus [P5]. But for some experts, transparency is above all a question of understanding the interests of others. One way of achieving this would be to display experts' affiliations.

Indeed, during meetings, there is no obligation to display one's employer, as experts are supposed to represent a form of neutrality with regard to their home institution[62]. As a result, most experts do not. The only obligation at plenary meetings is to display their surname, first name and country. This creates tensions because some experts fear that industry interests are interfering with standardisation work, but they are not allowed to point fingers [P5]. Indeed, according to the CEN-CENELEC Code of conduct, *"experts and observers are committed to revealing neither the identity nor the affiliation of other participants when using information received but not included in official minutes without prior consent"* (CEN-CENELEC, 2018). So, according to [P5], when an entire group opposes a proposal and all the members of that group belong to the same parent company, this behaviour, which should raise questions, is ignored because their affiliation is not public. No one can then contest, as otherwise they would risk being in breach of the code of conduct and being asked to leave.

In working group meetings, the rules are less strict and convenors can make up their own. In some groups, for instance, the leadership has asked experts to display their country and employer. This practice has then become more widespread over the last two years of JTC 21's existence and most groups now use it as a rule [P5]. Displaying affiliations at meetings also allows them to be recorded in the minutes, making it easier to trace lobbying activities. However, this does not solve the problem of consultants who will simply display their consulting firm and not their client, thus still concealing their true interests, but it is a first step towards more transparency.

[62] There is no obligation to display one's employer at European and international level. However, in some NSBs, such as in France, affiliations must be displayed, at least at national meetings (AFNOR, 2016, p.5). However, in our experience, this is rarely the case in practice.

Enable recognition of individual work

Today, academic researchers find it difficult to get involved in standardisation. Indeed, the entire standardisation system is not tailored to them, but rather to the industry. There are several reasons why academics are not more involved in standardisation: there is little funding available, there is no recognition of individual work, and standards are not very technical and academics do not want to get involved in political debates[63]. The European Commission itself has acknowledged that changes are necessary to attract more academics. In its recommendations on standardisation (European Commission, 2023b), the Commission states: *"The researchers should receive recognition for career development related to their contribution to standardisation"*. There is however no mention on what the solution to recognise their contribution should be. In addition, the Commission places the burden of finding these solutions on research institutes. [P5], themselves an academic researcher, suggests that the names of the contributors could appear directly on the standards, as they would for an academic article. But this solution is far from perfect.

First of all, this does not fully represent the work invested. For [P5], developing a standard can take up to five years. Summing this up in one line on a standard will not do the authors justice, but it would be a start. However, some experts are strongly opposed to this. Indeed, standards are supposed to be representative of everyone and should not single out individuals. *"I think standards are supposed to reflect the state of the art and best practice in a sector, not personal opinions. [...] So, no, I think it is right to remove any personification of standards. They are industry documents, representing a sector"**, [P9] explains. At a time when the Commission is increasingly urging the inclusion of more academics in standardisation, the question arises as to whether the standardisation system should adapt to the academic world, or whether it is up to the academic world to adapt to the standardisation system.

8.6.6 Should we re-imagine the system?

Some of the experts we interviewed were more radical in the solutions they envisaged. They believe that the standardisation system needs to be completely overhauled if it is to function properly.

An obsolete structure in a globalised world

Experts who want to change the system often start by criticising the structure based on national bodies. To them, in a globalised world, this structure is obsolete. [P8] explains that although NSBs make their own rules about who can join and become a member, most of the time, if the person speak the language used during meetings, this is enough to register. *"You could imagine a rule requiring you to be a citizen of the country, or to live there x% of the time, but there is no such rule"**, says [P8]. [P11] explains that there are multilingual experts who are registered with several NSBs at the same time and who can contribute to the positions of different countries. This is particularly useful for experts who are employed by different companies, representing their interests in different NSBs [P14].

For [P8], this national system no longer makes sense now that countries' interests are aligned: *"Are the views of France and Germany really so different today? [...] In the post-war period, people did not think about Germany in the same way as they do today. Today, I think you have a convergence of points of view [...]. So this structure based on*

[63] See Section 8.6.2.

*delegations aims to balance points of view that will converge anyway"**. For [P9], the problem is that companies and other structures take advantage of this system today. *"The whole standardisation system was imagined at a time when transnational companies did not exist. And so everything is structured around the countries that have their own national industry. It is a very pyramidal structure, with lots of little pyramids placed alongside each other. And nobody at the time thought that one day there would be horizontal players in all the pyramids, in all the countries. And that is where the GAFA have incredible power now, because they can manipulate all delegations at once. Something that would normally be inconceivable"**, [P9] explains.

A growing discontent about the business model

The experts also criticise the functioning of these national bodies, whose business model is based on memberships and the sale of standards. [P11], an industry expert, complains: *"[NSBs] are sometimes exorbitantly expensive. [...] You pay a membership fee to work at international, national or European level, you send resources at your own expense, you send people travel or whatever, you create a standard, you are the main contributor, you can be an editor, [...] and once the standard is finished, is published, you have to buy it"**.

[P3] also believe that it is not right for stakeholders to have to pay to access standards. For [P3], the very closed model on which standards are based is inexplicable in a world where the trend is towards free access to information on the web. *"It is a system from another century, it does not work at all"**, [P3] says. However, despite the discontent, the standardisation system is doing well, with ISO's total revenue from membership fees and the sale of standards exceeding forty five million euros a year (ISO, n.d.a)[64].

But what scandalises some experts the most is that harmonised standards are produced and financed in the same way, so that they are not openly accessible [P3]. However, as [P11] acknowledges, this business model is now being called into question. This dissatisfaction with standardisation is reflected in recent case law, which have ruled that hENs, in particular, should be made public[65]. [P11] explains that some experts are in favour of "smart" open standards, although they admit that they do not know how such a system would work.

Rearranging groups

When asked what they would change in the system if they had the chance, some experts imagine alternative solutions, where forces could be distributed to balance interests, instead of separating by countries. *"Perhaps the balance needs to be found [...] between foreign-based megacorporations [...], small businesses and academics. Perhaps, in the end, delegations should be made on this basis"**, imagines [P8]. [P9] envisions a system where transnational companies would be separated from the NSBs and have their own group: *"Their voice should be considered as a country. So they become their own country and no longer have the right to intervene in national bodies. In other words, there is a country called [Company name A], and a country called [Company name B], which is separate"**. This would force these companies to have only one voice and not to control the voices of the several countries.

But such a system could have multiple drawbacks. SMEs with an international presence could also be rejected from NSBs, or they could be forced to choose just one NSB in which to express their interests [P9]. Getting rid of the big compa-

[64] It should be noted that ISO's financial figures are given in Swiss francs. Today, 1 CHF is equivalent to 1.05 EUR.

[65] We examine this decision of the Court of Justice of the European Union in Chapter 7, Section 7.3.

nies would also mean that the country delegations would not have the expertise that these groups usually bring to the table. *"They've got experience, they know the processes, they can get things done and share knowledge"**, admits [P9]. Industry experts are obviously strongly opposed to this solution. However, [P11] admits that they would be happy to get rid of the "hegemony" of the NSBs, which are a major source of expenditure for companies, and to find another structure to participate in standardisation.

On the contrary, some would prefer to give more power to the NSBs, who currently have no say in the selection of experts. A solution was proposed by [P15], whereby NSBs would select experts independently of companies – and who pays the most, in order to have greater control over their national experts. Funding of the NSB would be independent of company participation and paid for by each company, which could then propose its expert to the NSB in a sort of recruitment process with no guarantee that its expert would be selected. This would oblige the experts not to take too strict a stance, or they risk not being re-elected. However, it is not clear how this "all companies pay" framework would be implemented in practice.

A system difficult to change

During the interviews, the experts were asked not to restrict themselves and to think of solutions as if they had a magic wand enabling them to change what they wanted instantly. However, things got more complicated when we asked them whether the solutions they proposed could really be implemented.

First of all, processes often seem untouchable because they are deeply rooted in the functioning of the standardisation system. This is because the rules are inherited from the parent bodies. For example, SC 42 inherits rules from JTC 1, which inherits them from ISO and IEC. So when a rule exists, it means that it has been validated by the entire chain of command. *"It is not easy to question something or to find a degree of freedom"**, reckons [P9]. Similarly, experts do not believe that it is possible today to modify the structure based on NSBs and organise the delegations differently. *"The system is already so locked down, so padlocked with rules, so no, something like that is impossible. We would have to create a new standardisation thing, completely new and different"**, says [P9].

When ask why they do not simply leave the NSBs, [P11] replies: *"To go where?"**. Indeed, the system is based on the cooperation between the NSBs and the SDOs, which means that it is impossible to participate to standards development without them. For [P11], the problem of the NSBs' business model is in fact inherited from ISO. These bodies, at national, European or international level, have a monopoly on standardisation. For [P8], the system does not change because stakeholders do not question it. There is a lack of organisation between them to propose an alternative solution. For [P15], the solution can only come from big companies, who have the economic power to demand change, or from the political side, which has the means to implement that change. But as [P4] points out, the problem is also monetary. Companies fund NSBs, so NSBs cannot afford to oppose them. NSBs therefore know that there is a problem with transnational companies strangling the whole system, but they cannot act for fear of losing their rent.

Not everyone wants this system to change

Only one expert told us he would not change anything [P12]. For some others, standardisation could be much better with just a few minimal changes. To our great surprise, none of the experts proposed simplifying the standardisation processes, such as the procedures for adopting or drafting standards. However, one of them mentioned the possibility of adding an appeals procedure to make it easier to challenge a decision.

Experts have different answers to the problems of standardisation today and how to solve them depending on their interest groups and how long they have been in standardisation. Indeed, new experts have only recently discovered this world and are indignant about the way it really works, while "governance" experts, usually from large companies, who are used to the existing dynamics, have no interest in changing them. They are therefore firmly opposed to the changes proposed by other experts, whether they be minor changes or, even more so, the total reconstruction of a new structure.

8.7 Discussion

8.7.1 An important world that will shape AI regulation

A study in the heat of standard making

CEN-CENELEC is working on AI standards since the end of 2021, and was asked by the Commission to draft the harmonised standards that will support the AI Act in 2022[66]. For now, CEN-CENELEC has adopted a few ISO standards on AI but has yet to develops its own, notably the harmonised standards. However, in the standardisation request, the Commission sets the deadline for these standards at 30 April 2025. We are therefore conducting this study at an exciting time, when it is possible to witness the birth of AI regulation, and at an exciting place, at the heart of the European standard making process.

But it is also a time where the stakes are high and tensions between stakeholders are therefore at their highest. We witnessed these tensions when, during the interviews, we heard different versions of the same stories told by different experts with different points of views. We also felt like although each expert was happy to talk to us, they did it mainly to defend their positions. Whatever the stakeholder groups, their intention was probably to use us, the interviewers, as a platform to carry a certain type of discourse. And the discourses we heard were often very committed, especially from stakeholders who are not part of large companies and who therefore find it harder to make their voices heard. This behavior is all the more normal as now is the time to make things happen, to make one's opinion heard and to influence standards in one way or another.

In the heat of standard making, power dynamics are all the more interesting to analyse. A discussion on which criteria to include in a standard may hide a wider battle between companies, where each stakeholder strive to impose its interests. This is understandable, as the outcome of standardisation will have a major impact on the economic benefits for companies. Having power over standards therefore means having the power to regulate the market. Even without the European system based on the NLF, the economic power held by these documents, such as ISO stan-

[66] An initial draft standardisation request was published at the time (European Commission, 2022f).

dards, is unprecedented due to their strong reputation and widespread adoption throughout the world. It enables the organisations which control standards to steer the market in their favour and dominate the competition. In the case of harmonised standards, the stakes shift, as the legal effects they produce enable stakeholders to influence not only the market, but on the law itself.

But companies are not the only ones interested. Harmonised standards are also a formidable means for civil society organisations to assert the interests of the group they represent. Power is acquired by contributing more than others, by creating coalitions and pushing votes in a certain direction or by obtaining leadership roles, such as editor, convenor or head of delegation. Which country leads the discussions, or which company steer decisions, therefore becomes a question of governance, at a time when Europe fights to remain sovereign, but is impacted by transatlantic and Asian influences. As geopolitical and economic dynamics continue to permeate AI, the concept of "standards war" (Shapiro and Varian, 1999; Stango, 2004), formulated in the 90s, is more relevant than ever.

Diverse experience of standardisation but one common goal: publishing AI standards

For this study, we interviewed a broad sample of standardisation experts from different countries, sectors, areas of expertise and levels of experience. While we note that the experts are fairly consistent in their descriptions of standardisation work and processes, the criticisms they raise are different, and the solutions they imagine even more so. We have identified several areas of great divergence: their personal difficulties – if any – in working in standardisation, what standards should contain according to them, what they think are the problems with standardisation today, and what solutions they are considering. Indeed, standardisation is multifaceted by definition, due to the variety of disciplinary fields and categories of stakeholders who are represented and who experience standardisation in different ways.

Competing interests create tensions that slow down discussions and delay standards. However, as the aim of standards committees is to publish standards, experts must inevitably find common ground to reach agreement. Ultimately, it is all about about aligning interests. But the task is colossal. While the BigTech often plays the role of the villain in the stories of some of our experts, for experts from these companies, it is necessary to get away from this Manichean representation. What is certain is that the truth is much more complicated than a two-sided story, with the big companies on one side and the others on the other. Even the academic experts we interviewed recognise that certain interests can sometimes conflict, because the big companies themselves cannot always be on the same side. For example, when geopolitical issues come into play, Chinese companies can team up with academic experts to block proposals from American companies. The map of the standardisation space is difficult to draw and is constantly changing according to which stakeholders decide to get involved and the positions they adopt. In this ever changing space, one constant remains: for the experts to reach an agreement, they have to work together. Several experts therefore stressed the need for mutual respect and understanding.

8.7.2 A small world seeking to open its doors

"entre-soi" in standardisation

AI standardisation is a very small world. The number of people actively involved is very low – according to some of the experts we interviewed, around fifteen people for all of the European standardisation initiatives – compared to the several hundreds of people registered in WG 4 alone, who are not active but simply monitoring the progress of standardisation work. We tried to interview as many "active" people as possible, but it was difficult to estimate their level of participation. This small number of people means that everyone knows everyone else in the world of AI standardisation. Experts regularly suggested names of other experts we could interview next, or told us they had heard we had interviewed a certain person because they had spoken to them. While this represents a challenge in itself in terms of guaranteeing anonymity, it also means that the daily lives of experts are influenced by a handful of stakeholders they see everyday.

While we have seen that experts find it hard to get into standardisation and learn the rules, and that many people get bored and only come for a few meetings before disappearing, those who manage to survive become masters of this field and even sell this expertise outside. As [P11] told us, standardisation is *"a job where experience is everything"**. It is also a closed world where it is hard to lie about one's involvement and expertise. An expert's reputation is based on the fact that they devote time to a specific subject and make themselves known.

Since negotiations are central, experts need to know each others. [P12] remembers that when they started out, they were too direct in their discussions with other members, too firm in their positions. Even more than diplomacy, this world is therefore about human relations. Experts make allies, friends and enemies, depending on how they behave towards each other. Tensions sometimes arise because of personality conflicts. This makes the world of AI standardisation even more complicated to decipher, because it is not always about group dynamics, but about individuals and their personal opinions.

A little-known system

There is a significant gap between the importance of standardisation and the interest shown in it by academic literature, the scientific world and the general public as a whole. There were a number of economic studies in the 1980s and 2000s, and then, with the NLF in Europe, legal literature also took up the subject[67]. But standardisation remains a very niche subject, where the dynamics of standardisation and the actual content that is produced are rarely analysed. Therefore, computer scientists and engineers know little about the work of standards committees. *"It should be part of our state of the art"**, says [P8]. However, access to standards is expensive and creates a gap between the world of standardisation and the world of science. Similarly, the general public has rarely heard of it, with the possible exception of ISO. But even when they know the names, they find it hard to understand what is really going on behind the walls and why this activity exists in the first place.

As [P11] explains, there is a bad buzz about standards today, because they are seen as a brake on innovation. On the contrary, their positive effects are not often mentioned. [P6] points out that the role of standards is first and foremost interop-

[67] See our analysis of the literature on standardisation in Chapter 4, Section 4.4.

erability. Standards enable many systems to function and simplify people's lives. Thanks to standards, phones can communicate, printers can use the same paper format, EU citizens can travel from one country to another on the same railway without having to change trains, they can use a USB stick on different computers, have a single charger for several electronic devices, use their credit card in cash dispensers in different countries, and so on. Standards therefore play an important role in the daily lives of citizens and, with the NLF, they have an additional role in supporting European law. For these reasons, they constitute an important object of study which deserves more attention.

Removing barriers

If we had to highlight one thing that we noticed during these interviews, it would be the experts' willingness to talk, their openness and their kindness. With the exception of one person, all our interview requests were met with a positive response and we really felt that the interviewees were happy to share their expertise and help us understand their world a little better. Given the stakes involved, more and more people are trying to take part in the discussions, but there are still a number of obstacles to overcome, including heavy procedures, workloads and lack of recognition for individual work and funding, which mainly disadvantages academics. Thanks to strong networking and determination, some national delegations are fighting to be as diverse as possible. But there are still efforts to be done to make standardisation as accessible as possible. Having an even more diverse standardisation space would, first, enable more interests to be represented, balancing out the super-power that certain groups currently wield, but it would also lighten the workload of current experts.

Alongside the experts' proposal to improve the standardisation system, there is also the complementary task of bringing this world to a wider audience. For [P11], this can be achieved through education. While there are already ISO-developed programmes for communicating in schools and universities[68], this movement is still struggling to reach Europe. But it is also a cultural issue. [P11] explains that in Germany, it is not unusual to learn about standardisation at school. France, on the other hand, is lagging behind, although some companies are beginning to set up courses and partnerships with universities. We believe that there is a real need for information on standardisation, both at school and in the professional world. However, this education and awareness-raising cannot be left to a few industry actors. Although these players have a necessary perspective on this world, education on standardisation is not complete without the point of view of academics or civil society stakeholders, who must also be involved in these awareness-raising initiatives.

8.8 Limitations and future work

Since CEN-CENELEC JTC 21 is a European group, we tried to interview representatives from different countries. Yet, due to our connections with the French delegation, we interviewed a majority of experts registered through AFNOR: 10 out of 16[69]. We must also bear in mind that the data we have collected comes solely from WG 4 on Foundational and societal aspects of AI and may therefore not be representative of CEN-CENELEC as a whole. However, as the future main har-

[68] Such as young professionals programmes (ISO, n.d.j).

[69] Or 62.5% of French representatives.

monised standard is being developed in this working group[70], we have reasons to believe that this is the most comprehensive working group, with the largest number of people registered.

The aim of this research work is, among other things, to highlight the systemic problems which exist in the current standardisation system, in order to try to find solutions. As a result, we work on a meso scale, looking at broad trends rather than individual trajectories and opinions. However, some of the problems raised by the study directly concern certain companies and individuals, whom some experts mentioned directly in the interviews. Nonetheless, in this chapter and in the thesis as a whole, we have chosen not to reveal any names of companies or organisations which are seen by some as problematic. We take this decision knowing that it creates an additional limitation to our work: that of speaking in general terms without directly naming those responsible. We stand by this choice because we believe that it is not an obstacle to understanding the general power dynamics that we are trying to highlight with this work. If the standardisation system is to be improved, we believe it will be by changing the system and the rules themselves, not the players.

Finally, the main contribution of this work is to shed light on the emerging field of AI standardisation and its ecosystem at national, European and international level. Our sociological contribution is therefore limited, as work on these organisations is only just beginning. Future work could focus on a better typology of stakeholders, their professional trajectories and their interactions at the micro scale.

8.9 Conclusion

In this chapter, we presented the results of fieldwork at CEN-CENELEC, where we collected data and interviewed 16 standards experts currently working on AI-related standards.

This study enabled us to gain a better understanding of the variety of expert profiles, because although the industry plays an important role in standardisation organisations, a large number of members from civil society are also involved. Stakeholders find a wide range of interests in standardisation work, from the creation of a professional network to the gathering of experience, all of which seem to be profitable reasons for engaging in this work, even if the prospects are sometimes long in coming.

These testimonies also highlight the gap between the standardisation experiences of the various stakeholders. Experts mostly agree on the fact that they are understaffed. But experts which begin standardisation work with AI also complain about the overly lengthy and complex processes imposed by the organisation. Furthermore, the standardisation system is not adapted to academic researchers as there is little funding available and no recognition of individual work. Yet they are essential if Europe is to realise its ambition of having multi-stakeholder representation in AI standardisation. According to experts outside of large companies, the distribution of experts in national delegations further gives disproportionate power to transnational companies, which can afford to have representatives in several countries. This phenomenon is exacerbated by the financial and human cost of standardisation work, which only allows large companies to invest manpower in standardisation.

Regarding the content of standards, CEN-CENELEC has undertaken to develop

[70] Namely, the AI trustworthiness framework. For more information, see Section 8.5.3.

a single standard to cover all the requirements of the AI Act: the "AI trustworthiness" standard. However, the amount of work to be done and the potential overlap with other standards complicates the task and means that this standard, and others like it, are falling behind schedule. Subsequently, there is a growing pressure to adopt existing international standards, particularly in the area of risk management. This pressure is exerted mainly by experts from large companies who also participate in international committees. But the adoption at European level of standards developed outside Europe raises political and sovereignty issues.

The solutions proposed by the experts range from small changes to the standardisation system, such as greater transparency and recognition of individual work, to a complete overhaul of the national-based system to replace it with a structure based on group interests. While radical change is unlikely at this stage, if the standards are not satisfactory, the European Commission may develop alternative frameworks, such as common specifications.

While the criticisms of standardisation which are resurfacing in the discourse of our experts are not entirely new, this is the first time, to the best of our knowledge, that standardisation experts have been given the floor to address these issues themselves and propose solutions. This study is therefore aimed both at the general public, seeking to understand standardisation work, and at the standardisation experts themselves, in order to address unspoken issues and initiate discussion between different stakeholders groups. It could also help the European institutions to see the see the intricacies of the standardisation system and to take the appropriate actions to remain in control.

Standardisation is a highly diplomatic ecosystem where different entities try to impose their own vision, based on their economic and political interests. If AI is to be governed by technical standards, we need to shed light on the work produced within these private organisations and ask ourselves how we can best accompany them to ensure that AI standards will indeed support EU values and interests.

Acknowledgements

We would like to thank all the participants in this study for taking the time to talk to us. Each interview offered a new perspective and insight into the multi-faceted work of standardisation. We hope this study reflects and honours your different experiences and that this chapter, and the thesis as a whole, will contribute to a better understanding of the world of standardisation.

DISCUSSION AND CONCLUSION

Chapter 9
Discussion: regulating AI through vertical sectors

Where does this chapter come from?

This chapter is an original contribution from this thesis. It continues the various topics of discussion begun in the previous chapters and draws on their conclusions to state personal opinions and recommendations for the regulation of AI in Europe. It can be seen as a position chapter, describing the authors' views on the issues that have been raised throughout this thesis.

It should be noted, however, that considering the number of different topics covered in this thesis, the areas for discussion are also very broad. Each of the six sections of this chapter therefore addresses a specific point and readers are encouraged to go directly to the parts that interest them.

This chapter was also released online as a standalone work:

> Gornet, M. (2024d). Too broad to handle: can we "fix" harmonised standards on artificial intelligence by focusing on vertical sectors? `https://hal.science/hal-04785208`

However, it has since then been updated and completed, especially with a new Section 9.7.

Caveat
For this chapter, we have used some of the material collected during the interviews with CEN-CENELEC standardisation experts that we conducted for Chapter 8. We will not repeat here how the information was collected and processed. Interested readers can take a look at our methodology in Chapter 8, Section 8.3.2.

Chapter's abstract

The European approach to regulating AI has relied on three main regulatory mechanisms: ethics charters, the AI Act and technical standards. Europe has based this approach on concepts such as "trustworthiness" or "risk", navigating a conceptual space where the ethical, legal and technical fields clash. The origins of this approach in ethics charters, which usually focus on broad principles, have led to the dissemination in the AI Act and in standards of a very general discourse about AI, which rarely goes into technical detail, and with elements that are unimplementable as is. Additionally to this broadness of principles and requirements, the European

discourse on AI, whether in ethics charters, the AI Act or standards, has also remained very horizontal. While the AI Act classifies high-risk systems according to their sector of use, the obligations applicable to them are the same regardless. This poses a problem for standards, which are forced to remain at a high level, as the technical requirements are too difficult to define without contextual elements. We therefore propose to refocus standards on vertical sectors and technologies, allowing them to define stricter requirements.

9.1 *Introduction*

AI regulation has been a topic of interest for the last fifteen years, with the publication of various ethics charters, legal texts and technical requirements around the world. This movement reached Europe in the late 2010s, with the publication of the High Level Expert Group (HLEG) guidelines (HLEG, 2019a) and, more recently, of the AI Act (European Parliament and Council, 2024b), the first mandatory framework for AI in the world. Since then, Europe has refined its approach based on "trust", a philosophical concept that has become a compliance tool. Indeed, in the AI Act, providers of high-risk AI systems will have to respect a number of obligations, often by testing the technical properties of their systems (Art. 9 to 15). The details for the practical implementation of these requirements are not defined in the AI Act but in technical standards. Subsequent efforts have therefore focused on developing a certain type of standards, harmonised standards (hENs), which are technical documents drafted by European Standardisation Organisations (ESOs), and in particular by CEN-CENELEC[1] JTC[2] 21 in the case of AI systems, at the request of the European Commission. These hENs, if they are cited in the Official Journal of the European Union (OJEU), will be granted legal effects (Gornet and Maxwell, 2024). This includes the presumption of conformity with the law, when a provider is compliant with the respective hENs. This approach to product safety is common in Europe since the 80s, and is called the New Legislative Framework (NLF).

The three-step approach to AI regulation in Europe – with charters, law and standards – has, however, generated much discussion in the literature because of its distinctive features. Indeed, the AI Act mixes rights and risks and therefore asks standards to address fundamental rights issues for which they are not equipped (Gornet and Maxwell, 2024). Further discussions with standardisation experts also revealed that the horizontal and cross-cutting approach of AI standards, inherited from ethical charters and the AI Act, is difficult to implement in practice[3]. What are the problems raised by the European approach to AI regulation and can we suggest improvements?

This chapter acts as a position paper, where we rely both on the analysis of previous literature and European institutions documents, as well as on our previous study in Chapter 8, where we interviewed 16 experts from CEN-CENELEC JTC 21. Their names have been anonymised and have been replaced by P1 to P16. When their testimonies have been translated from a language other than English, the quote is marked by an asterisk (*). For more information about the interview protocol and the experts' background, see Chapter 8.

We begin in Section 9.2 by examining ethics charters, the AI Act and technical standards, how they each draw inspiration from the other two, and we show that

[1] European Committee for Standardisation (From French: Comité Européen de Normalisation) and European Committee for Electrotechnical Standardisation (From French: Comité Européen de Normalisation Électrotechnique).

[2] Joint Technical Committee.

[3] See Chapter 8, Section 8.5.

this is not necessarily beneficial to the regulation process. In Section 9.3, we look at the content of standards and show that, although they were originally intended to define specific technical requirements, with AI, they tend to focus on governance and normative issues and we question whether this is the right way forward. In Section 9.4, we therefore show that AI standards could mark the beginning of a fourth wave of standardisation, with new topics being addressed, by a wide number of organisations and with a diversification of experts' profiles. In Section 9.5, we recall the problems that the standardisation system is facing, from the restricted access to standards, the influence of international voices and the private sector, to their general lack of legitimacy. We show that this is mainly due to a shift in standardisation, towards the inclusion of more diverse stakeholders at the request of the Commission, the horizontality of standards and the strict deadlines to which they are subjected. We therefore propose in Section 9.6 an alternative solution to the current AI standardisation strategy: refocusing standards on vertical sectors, where they could define more specific technical requirements and appeal to a more diverse audience. However, this will not come without some cost, given that there is no more time before the deadline for harmonised standards, and standards for the AI Act are expected to be horizontal. Additionally, it could be difficult for cross-sector providers to navigate a myriad of vertical standards. Finally, in Section 9.7, we conclude on the pros and cons of the AI Act approach and the legacy it leaves behind.

9.2 *Ethics, law and standards constantly overlap*

9.2.1 *Technical and legal influences on ethics: a loss of meaning*

Today, with AI in particular, ethics is wrongly applied to a technological object. In its traditional sense, ethics is associated with moral behaviour (Merriam-Webster Dictionary, n.d.g) and sets of principles intended to guide a person's actions. It therefore applies to reasoning or a moral being and, like trust, cannot become a characteristic of AI.

Nevertheless, AI ethics charters have become accustomed to defining these principles in the form of checklists, which resemble more good development practices than virtuous behaviour. This technicisation of ethics is observed by Hagendorff (2020), who explains that technical elements are conceptualised in AI ethics charters as *"isolated entities that can be optimised by experts so as to find technical solutions for technical problems"*. There is therefore a risk of diversion from the traditional conception of ethics[4].

In addition, this approach to ethics as a set of technical properties is also characteristic of what Hunyadi (2015) calls "Small Ethics"[5]. "Small Ethics" is defined by the author as *"the liberal ethic of individual rights and freedoms"* (Hunyadi, 2018), it is small not because it is not important, but because it is centered around the individual rather than having a global perspective. In the context of new technologies, "Small Ethics" is part of a *"fait accompli"* logic (Tessier, 2022), where individuals must adhere to these ethics principles and the lifestyle they convey, thereby restricting their freedoms. This is also called *"ethics of carefulness"*, or *"checklist ethics"* (Bolte et al., 2022). This *"checkbox culture"* therefore runs the risk of taking the responsability away from AI practitioners, who simply follow the list without taking into account the actual harm caused by their systems (Balayn et al., 2023).

[4] What (Tessier, 2022) calls, in French, *"le dévoiement de l'éthique"*.

[5] From French, *"Petite éthique"*, although the translation "Small Ethics" is used by the author himself (Hunyadi, 2018).

As part of the EU's approach to AI regulation, ethics guidelines, such as those from the HLEG, were also a preparatory step to mandatory legislation, preparing concepts such as "trustworthiness" and "risk", which would be reinforced in the AI Act. However, ethics frameworks are meant to complement legislation, not prepare it. Ethics should not be solely a preparatory stage for mandatory legislation, nor should ethics be constrained by law. On the contrary, it should try to go further than the law, because respect of the law is the minimum requirement in a democratic society. The two approaches complement each others and should be developped in parallel.

9.2.2 *Ethical and technical influences on law: the European "trustworthiness" strategy*

AI projects and initiatives, whether research articles, policy documents or industry communications, tend to use catchy keywords to surf on the hype of AI ethics (Probasco et al., 2023). Different spheres of AI stakeholders are publishing their own charters or policy documents, using a rather convergent discourse, but which still conceals significant divergences, specifically on the vocabulary that is used. A potential explanation to this mix of discourses is the influence of both the industry and the civil society on governmental and supra governmental entities which become arenas to push one's interests[6].

Since the HLEG guidelines in 2019, the European strategy for regulating AI has been based on a notion of "AI trustworthiness". European institutions justify this choice by explaining that "trust" is considered to be the *"bed-rock of societies"* (HLEG, 2019a, p.4). For Laux et al. (2023), citizens are encouraged to trust AI so that they can use it more and businesses and governments can reap its economic benefits. Indeed, studies have shown that greater trust often translates into positive economic development (Bjørnskov, 2017). There is therefore a deliberate confusion, in the AI Act, between "trustworthiness" and "acceptability" (Laux et al., 2023).

The HLEG defines "trustworthiness" as a concept based on three pillars: ethical AI, lawful AI and robust AI. But this definition is far from being a given in the literature on trust. For Nickel et al. (2010), there are two types of trust: the pure *"rational-choice"* notion of trust, which cannot be differentiated from the engineering notion of "reliability", and the *"motivation-attributing"* notion of trust, which require the trustor to attribute some motivations to the trustee with regards to their values and interests. In that last sense, AI cannot be "trustworthy" as it is not a human agent, and has no moral or motivations. Ryan (2020) adds that AI does not have the capacity to be trusted because it cannot be held responsible for its actions. For Ryan (2020), speaking of "trustworthy AI" is therefore a dangerous anthropomorphism, which wrongly attributes moral and emotional capacity to technology. But some authors still think an adaptation of the notion of trust is possible to fit the AI context. Rieder et al. (2021) proposes that an AI system could be considered "trustworthy" when it is responsive to the interests of the human trustors. As such, it does not need to have its own motivations, but simply to know those of the human party. Yet, despite a few exceptions, most of the literature agrees that AI cannot be "trustable", although it can be "reliable" (Laux et al., 2023).

In the context of the AI Act and AI standards, the standardisation experts we interviewed for this thesis give a definition of "trustworthiness" that is indeed closer to that of "reliability". *"We are working on 'trustworthiness' according to the European*

[6] See Chapter 2.

*Community's approach, which is a market approach. The market approach to 'trustworthiness' allows for possibilities based on certain characteristics"**, explains [P2]. These characteristics are the technical requirements that will be defined in standards to precise the essential requirements of the AI Act. This definition is coherent with the one given by international standards on AI where trustworthiness is defined as *"the ability to meet stakeholders expectations in a verifiable way"*[7] (ISO/IEC, 2020). On the contrary, reliability is defined as the *"property of consistent intended behaviour and results"* (ISO/IEC, 2020). While the two are clearly different according to standards, there are still inconsistencies that are revealed by translation into different languages. For instance, the "overview of trustworthiness in AI" proposed in ISO/IEC 24028 (2020) is translated in French by something much closer to "overview of reliability"[8].

We are therefore witnessing a semantic shift whereby a discourse on "ethics" has been transformed into a discourse on "trustworthiness", which, in the HLEG guidelines, is supposed to encompass ethics, law and technical properties, but which is gradually being understood and reduced to the latter dimension in the AI Act and in standards. This shift is justified by the German Institute for Standardisation (DIN) because, in their view, ethics refers to rational beings and trustworthiness could refer to organisations and technical systems (DIN and DKE, 2022). However, the emphasis in Europe on trustworthiness rather than reliability appears to be part of a wider communication strategy to make AI more acceptable. Acceptability is therefore seen as an objective, and "trustworthiness" as a means of achieving that objective. The approach is similar for "risk" where, apart from a few unacceptable ones, many risks are considered, in the AI Act, acceptable under the right conditions. As such, the EU's approach to AI regulation with, in particular, the AI Act, which is based on a specific version of trustworthiness, advertise its ethical origins, but is actually more influenced by technical notions of reliability.

9.2.3 *Legal and ethical influences on standards: standards are becoming more than industry documents*

Technical standards for AI are increasingly trying to address normative questions, which are usually reserved for ethics charters (Gornet and Maxwell, 2024). This trend is particularly apparent in Europe , because of the European Commission's desire to address the risk to fundamental rights in its AI Act. As a result, the ESOs have adopted a horizontal and high-level approach to AI standardisation, using the same keywords as ethics charters and the AI Act, with an upcoming standard called the "AI trustworthiness framework". The "AI trustworthiness framework" is intended to cover seven of the ten items present in the standardisation request (CEN-CENELEC, 2024b), which makes it very broad and horizontal, a characteristic quite uncommon for technical standards. This "trustworthiness" framework, like previous ethics charters, attempts to address technical criteria of reliability.

At the same time, standards have taken on a legislative role in Europe with the New Approach since the 80s. The combination of more normative issues and increased legal powers brings these frameworks closer to mandatory legislation. However, standards are advertised as documents drafted by so-called "experts", i.e. individuals who have technical and industrial expertise and who are capable of implementing these specifications in a business pipeline. This creates a disconnect between the nature of standards and their evolving role in European regulation.

[7] Stakeholder expectations include accountability, accuracy, availability, controllability, integrity, quality, reliability, resilience, robustness, safety, security, transparency, and usability.

[8] *"Examen d'ensemble de la fiabilité en matière d'intelligence artificielle"*, intead of *"Examen d'ensemble de la confiance"*.

As a result, standardisation organisations become increasingly politicised, raising concerns about their legitimacy.

9.2.4 *Interacting without blending*

Ethics has lost some of its meaning in the attempt to define more technical criteria in charters; the AI Act bases its entire strategy on "trustworthiness", inspired by ethical-technical frameworks; and standards, which were already considered legal acts under EU law, are becoming more normative with AI. As a result, the three spheres that the European AI regulation strategy is trying to address are actually quite intertwined, with increasingly blurred borders.

But with this blending, each of these frameworks actually loses its effectiveness and legitimacy, because the people who draw them up cannot be competent in all areas. According to Eliantonio and Medzmariashvili (2017): *"EU law, while embracing hybridity, has not yet found a way to come to terms with it"*. There are therefore calls to keep these spheres more separate. The European Parliament has notably stated, several years, ago that *"it is of the utmost importance to draw a clear line between legislation and standardisation in order to avoid any misinterpretation with regard to the objectives of the law and the desired level of protection"* (European Parliament, 2010). Similarly, the mix of technical standards and fundamental rights satisfies neither the standardisation experts nor civil society associations[9].

Regulatory frameworks that try to do too much run the risk of doing nothing properly. We therefore recommend avoiding this mix of approaches as much as possible and focusing each framework on what it does best: ethical charters to guide reflection and action, the law to protect rights, and standards to define specific technical properties. These frameworks should still interact, but all the while having their own separate scope.

9.3 *Technical standards are not that technical*

Standards are becoming less and less technical. This is due to two trends: (i) making standards more about management processes than about the products themselves, what we have called "governance" standards[10], and (ii) incorporating ethical, social and even fundamental rights issues into standards.

9.3.1 *Addressing governance versus technical properties*

Standards for AI fell into two main categories: either "governance" or "information" standards[11]. Governance standards concern the processes put in place by companies to produce AI systems, while information standards define technical terms and properties associated with AI systems. However, contrary to other fields, there is no "performance" standards for AI at the moment[12], i.e. standards setting minimum thresholds for certain properties of AI systems. These different types of standards can help to regulate AI in various ways.

Governance standards are flexible

Governance standards include quality management and risk management standards. They are more concerned with what companies can do and what mechanisms they can put in place to prevent something from going wrong, than with

[9] See Section 9.3.2.

[10] See Chapter 5, Section 5.5.1.

[11] According to our own classification. For more information, see Chapter 5, Section 5.5.1. It is different from the usual separation between types of standards in the academic literature – see Chapter 4, Section 4.2.2, or that of the SDOs – see Chapter 4, Section 4.5.3, but we felt that this classification made more sense for our study, and in particular for the following question on whether standards should address ethics and fundamental rights.

[12] We examine the possibility of such standards in Section 9.3.2.

checking whether there is in fact an issue with the product. The purpose of governance standard is very different from that of product standards [13], so they cannot easily be compared. It is the most widespread type of standard in the modern era (Yates and Murphy, 2019), particularly for new technologies where agreement on the constantly evolving state of the art is virtually impossible (Choi, 2024). This trend is, however, fairly recent. It developed in the 1980s with quality management standards and software standards and is now becoming the main type of standard for AI[14].

Governance standards are notably criticised for not being strict enough and for leaving companies too much room to manoeuvre. Indeed, companies often do not need to change their existing processes, and compliance to the standard is used as a justification for bad practices. As [P1] puts it: *"[auditors] simply check that the company is achieving the objectives it has set itself"**. This lack of control can be dangerous if it is set up as a mark of quality used as a marketing tool by companies, but it is even more dangerous when it becomes the only safeguard required by law. These standards have also been criticised for not being adapted to SMEs or startups. For example, the ISO/IEC 42001 (2023b) standard, on risk management for AI, deals with the organisation of the company and requires the presence of several teams – legal, technical, etc. – a structure that a small startup simply cannot afford. Rendering the standard almost compulsory by harmonising it could therefore kill off small businesses. In that regard, the European Commission, which works closely with standardisation organisations, including CEN-CENELEC, has already stated that ISO/IEC 42001 was not enough for supporting the AI Act risk-based requirements. Indeed, the AI Act requires providers of high-risk AI systems to have a risk management framework in place, a framework that will be standardised by future hENs. ISO/IEC 42001 has therefore been adopted by CEN-CENELEC and has become a European standard, but is not in the process to become a harmonised standard to support the AI Act.

But just because ISO/IEC 42001 is not sufficient to ensure a sufficient level of protection for Europe does not mean that governance standards are useless. Particularly in the context of the AI Act, companies will need to put in place a risk management framework for high-risk systems. These frameworks will need to be standardised as the AI Act relies heavily on standards to implement them. Without effective risk management standards, the legal text would risk being unenforceable. Yet Europe needs to be cautious about the level of requirements that these standards convey, particularly if they are to support the AI Act. In particular, they must be applicable to all companies and be sufficiently strict not to justify bad practices and enable effective supervision. Above all, they must not be used as the sole means of protection but they can complement more technical standards, such as information standards.

Information standards avoid technical details

Information standards are often more technical than governance standard. Even if they do not set thresholds, the technical definitions and benchmarking possibilities they offer can push companies to innovate and compete to achieve a higher level of protection. This protection includes security measures as well as measures relating to fundamental rights[15]. However, even when it comes to definitions and measurements, the question arises as to how far standards should go into technical

[13] Both information and performance standards.

[14] For more information, see Chapter 4, Section 4.4.1.

[15] See Section 9.3.2.

detail. Indeed, the standardisation of techniques that are not sufficiently mature has proved to be a considerable brake on innovation in several areas prior to AI. This is why some experts claim during standard development that a measure cannot be included in a standard, because although it is widely used in the field, its effectiveness has not been proven [P5]. But at the same time, not going into the technical details of the technology makes a standard virtually useless. This is why some standardisation experts are in favour of including more definitions and measurements in standards in general, at the risk of them becoming obsolete in a few years' time. Others, are more cautious about what they want to include, at the risk of having standards that are emptier and further removed from the current state of the art. The level of maturity of technical properties, tests or measurements that is sufficient to incorporated them into a standard is therefore relative and not an exact science.

At the same time, it is interesting to examine the dynamics of the pressures exerted by the various parties in the context of standardisation. BigTech companies are often in favour of including less technical elements in standards, either by developing more governance standards or by making definitions and measurements in information standards less technical.

There are many examples of how companies can undermine mandatory regulation by refusing to include technical elements, such as precise metrics and mathematical formula, in standards, preferring broad and sometimes circular definitions. Examples include a definition of transparency that does not refer to the notion of access to the system for audit purposes, or a definition of data representativeness which is unimplementable[16]. This imprecision in standards, even the most technical ones, is due to the fact that standards, as a form of regulation, limit companies in their practices while companies would prefer to be as free as possible. Market dynamics also have a major influence on these positions, particularly for companies that intend to bring to market new innovative products that are technically far removed from what is described in the standards.

9.3.2 *Addressing ethical, social and fundamental rights questions*

A push for more normative questions in standardisation

The question of the appropriate technical level of AI standards should be complemented by the question of whether and how ethical, social and fundamental rights issues can actually be standardised. Indeed, the European Commission, with the AI Act, pushes the standardisation ecosystem to integrate these issues into standards. Even before the AI Act, ethics was cited as a potential subject of interest for standardisation organisations (CEN-CENELEC, 2020; DIN and DKE, 2022). But with the AI Act, Europe has been propagating a dual discourse, which aims to protect fundamental rights while at the same time having a risk-based structure inspired by product safety regulations. This has led the ESOs to develop a variety of standards which deal to some extent with ethical and fundamental rights issues, such as the standard on AI "trustworthiness", and other initiatives, on the competencies of AI ethicists or how to conduct a fundamental rights impact assessment for instance[17]. But Europe is not alone in its journey to standardise ethical, social and fundamental rights questions. Other initiatives have been launched by ISO, IEEE[18] and other organisations[19].

[16] See Chapter 8, Section 8.6.4.

[17] See Chapter 7, Section 5.3.1.

[18] Institute of Electrical and Electronics Engineers.

[19] See Chapter 5, Section 5.3.4.

Not everyone agrees that these questions should be standardised

However, addressing ethical, social and fundamental rights questions in standards can be hard. Apart from the Commission, other European institutions are cautious about incorporating these issues into standardisation, with the European Parliament stating that standards should not address *"fundamental rights or socio-economic issues"* (European Parliament, 2023c). Some oppose these initiatives, even within standardisation bodies, claiming that ethics is not "tangible" enough to be standardised or that it is not the role of the ESOs to standardise fundamental rights[20]. This type of statement seems to be common in engineering task forces and standardisation groups, with experts claiming that they *"do not do politics"*, or that they are *"just engineer[s]"* (Cath-Speth, 2021). Similarly, among the experts we interviewed, those with a technical background admit that they are not really interested in ethics [P15]. In a recent study (Baeva et al., 2023), a standardisation expert working on AI explained: *"We cannot write into a standard what is good and what is bad. What we can write into it, however, is a description of the ethically relevant properties of the system, i.e., the degree of transparency, the degree of fairness, the degree of privacy, the degree of robustness, and the methods for measuring them."* This statement seems to be in line with the EU's approach to "trustworthiness", which is closer to a list of technical properties that guarantee "reliability" than to ethical considerations[21].

There does not even seem to be a consensus on what constitutes such ethical and fundamental rights standards. Some standards, such as CEN-CENELEC "competency framework for AI ethicists"[22], are clearly considered by all to be "ethical standards", but more technical standards, such as fairness standards, are not always present in everyone's mind. The "AI Trustworthiness framework", for instance, is not always considered by CEN-CENELEC experts as an "ethical" standard. We found during our interviews that experts seem to differentiate between "technical" standards and all the other standards, which deal with ethical, social or fundamental rights issues, but at a higher level. This last category is sometimes considered the least important, because the main objective of SDOs is generally to produce technical specifications, and the experts themselves sometimes feel less entitled to discuss these issues. Additionally to standardisation experts themselves, various civil society organisations that focus on defending citizens' rights have similarly called for standards to be limited to technical aspects (EDRi, 2022; Giovannini, 2021).

Standards are not neutral

However, we do not subscribe to this categorisation of standards between "technical" and "everything else". In our opinion, the boundary between what is ethical and what is technical is difficult to draw, resulting in standards that may make value judgements without the developers of such standards even realising it. Fairness standards are a good example: deciding on a threshold for a specific measure of (un)fairness can be understood as setting a threshold for an acceptable level of discrimination, given that residual biases will always be present. With regard to performance standards, the choice of a threshold is a normative judgement, even for safety standards, putting a quantitative measure on the right to safety and sometimes even on the right to life of individuals. Safety thresholds, however, are more accepted and more widespread in our society than thresholds for fundamental rights. This raises the question of what makes fundamental rights different than

[20] For more reactions on the standardisation of ethics and fundamental rights by standardisation experts, see Chapter 8, Section 8.5.1.

[21] See Section 9.2.2.

[22] For a complete list of JTC 21 published standards and standards under development, see (CEN-CENELEC, n.d.b,n).

traditional safety issues.

One of the main differences between safety standards and standards relating to ethics and fundamental rights is the level of subjectivity and cultural significance. For example, technologies that are accepted in some countries will not be accepted in others. In that regard, the EU has taken a firm stance by banning social scoring systems, which are already widely used in China, which accounts for 17% of the world's population. There is therefore no consensus among the world's population on what use of technology is "ethical" and respects fundamental rights. As with the choice of whether or not to use a technology, the choice of whether a risk to fundamental rights is acceptable or not is a normative and political choice that is codified in standards, directly questioning what we accept as a society.

But what differentiates the AI Act approach from that of standards is that the AI Act is a legal text which, as such, necessarily codifies certain aspects of ethical behaviour and defines the rights of individuals. The AI Act can set normative thresholds on what is acceptable or not because the EU institutions have the legitimacy to make these choices. On the contrary, standards are drawn up by private bodies whose work is supposed to remain technical. Despite their power within the EU due to their close collaboration with the European institutions, ESOs are not supposed to have a say in policy-making.

Finally, there is also the question of the maturity of the technology. If we accept a residual risk in the case of nuclear activity, this is both because the alternative would be to stop the activity altogether, which would otherwise benefit humanity, and because the "residual risk" is very low, the chances of a nuclear accident being of the order of one in billions. In the case of AI, it is questionable whether the use of AI systems is sufficiently beneficial to society to take significant risks to fundamental rights by using them. Additionally, the threshold that standards would have to set would be much higher. Indeed, the "residual risk" to fundamental rights is partly higher because AI technology is still in its expansion phase, where new advances are made every day, and because of the lack of causality that makes the behaviour of AI systems difficult to predict.

The danger of setting thresholds

In our analysis[23] we have not seen any AI standards aimed at setting thresholds for the acceptability of risks to fundamental rights. This is not to say that it is not a possibility. In the case of fairness, as metrics develop, we could imagine one day having a standard that would draw the line between what is an "acceptable" level of biases and what is not. This threshold could be decided following different rationales: the standard could either fix an "olympics" threshold that do not settle for less than the best possible solution in the state of the art, or a "filter" threshold that only fixes an arbitrary limit (Busch, 2011). Each of these types of threshold allows a different version of what is "acceptable"[24].

In the case of a filter threshold, on the one hand, it would push industrial players to improve their technology, at least up to a certain point. On the other hand, it risks setting a threshold which, once reached, will slow down companies in their quest for improvement. Additionally, setting a filter threshold today would run the risk of setting limits that would be obsolete in a few years' time. Even if standards can be updated, this constant lag behind technological advances risks causing a number of breaches to people's fundamental rights, whereas the state of the art would make

[23] See Chapter 6 and Chapter 5.

[24] For more information, see Chapter 3, Section 3.7.

it possible to avoid many incidents.

Yet, when thresholds are set for criteria that may affect fundamental rights, even when they are olympics thresholds, they still allow certain harms to occur and to be permitted. For example, an AI system may still be discriminatory even when the provider has applied the most modern techniques to avoid bias. This discrimination could have happened either because the state of the art will never allow all bias to be eliminated, or because the system is used in a way that creates bias and discrimination[25]. Unlike a case where AI regulation would be rights-based, under the current AI Act, system providers who would have applied this olympics threshold would be excused because they would have done everything they could to avoid discrimination. However, such a threshold would still be better than a filter threshold where companies would only have to tick the boxes set up by standards without thinking about the possible other harms their systems could cause.

Remaining cautious while standardising information

But even when standards do not try to set thresholds, addressing ethics and fundamental rights still remains difficult. In information standards, the definitions given and the metrics listed can have an influence on the tests carried out and therefore on the problems that are revealed, but above all on those that remain hidden. We have seen that the chosen measure of fairness, for example, can lead developers to conclude that the system is or is not biased. The conclusion on whether discriminatory practices have indeed taken place is therefore closely linked to this technical choice. Presenting developers with a list of metrics without forcing them to test them all would only allow them to choose the one that proves that their system is not discriminatory.

On the other hand, standards must necessarily define the technical state of the art, and not acknowledging the existence of these metrics in standards would simply sweep the problems of unfairness under the rug. Particularly for standards which will support the AI Act, the requirements that appear in these standards are likely to be the only ones implemented by companies, at least initially. As such, they need to be as technically precise and as complete as possible, without taking any decision for the developers. The final decision of which metrics to use and what test to carry should mostly be left out to the responsibility of the developers.

If AI standards confine themselves to defining either processes in governance standards, or technical definitions and metrics in information standards, the design choices made by the developer as to which test to perform will carry a normative value of what they consider to be a "sufficient" level of protection. It will then be up to the judges, in case of litigation, to decide whether these technical measures were appropriate or whether they infringed fundamental rights.

9.3.3 *A question of balance*

The degree of precision of the metrics, tests and procedures that standards should include to ensure compliance with the requirements of the AI Act depends first and foremost on whether policymakers want a fully ex ante approach to AI regulation.

Indeed, if the aim is to prevent most damage before systems are put on the market, the right approach should probably be to define as many things as possible in standards, including olympics thresholds. This would also put an end to a great

[25] For example, the feedback loops that are created when the output of an AI system is used as input.

deal of damage, by making the systems safer and more respectful of fundamental rights. However, this would be to the detriment of the EU's other objectives, namely the free movement of goods and the desire not to hinder innovation. It could also be a heavy burden for smaller companies which will suffer from compliance costs. Additionally, residual problems that have not been defined in these standards will be rendered non-existent and it will be even more difficult to prove that anything has gone wrong. In short, this would make it possible to avoid certain prejudices to the detriment of others, which are less measurable and therefore less defined in standards, but which would nevertheless have a significant impact on fundamental rights. The biases of facial recognition systems are a case in point: until the outrage caused by academic studies and false accusations in the late 2010s, providers and lay persons alike were unaware that these systems could discriminate against groups of people. Although some of these biases have since been dealt with, there is no telling how many problems like this one persist in AI systems that we are unaware of. Giving too much power to standards that define the current state of the art could prevent the discovery of new problems. But above all, even if they are known, these problems will not be addressed by companies because standards do not require them to be. Furthermore, there is no guarantee that, even if there is the political will to define very strict rules for companies, the ESOs, which are mostly run by these companies, will comply and develop very technical and restrictive standards.

On the contrary, standards now have to face up to the possibility of not defining too many technical requirements, firstly so as not to upset companies and jeopardise the European market, but above all to leave room for interpretability of what constitutes "sufficient" protection. Less stringent standards would have the disadvantage of leaving some damage unaddressed when systems are put on the market. But at the same time, leaving it up to companies to decide what they want to put in place to protect individuals makes these choices more open to scrutiny and ensures that developers take responsibility for their decision-making. This type of transparency and openness, where providers of AI systems would only have to adhere to the same harmonised practices of testing, without obligations on the results of these evaluations, is what Laux et al. (2024) call *"ethical disclosure by default"*. But simple disclosure without action may not be enough to counter all the dangers of AI systems.

Firstly, this approach, which is in line with the promotion of governance standards, is supported by some industry players as it would allow greater flexibility in the implementation of requirements, but it is also supported by some academics who believe that strong normative choices should not be made in standards but should be dealt with by the actors involved in conformity assessment (Tartaro, 2024; Gonzalez Torres and Ali-Vehmas, 2024). What this solution neglects to say is that, under the AI Act, conformity assessment will be carried out mainly by the AI system providers themselves, who are sometimes the same people who wrote the standards.

Secondly, an approach based on simple transparency runs the risk of simply having suppliers carrying out audits in accordance with the harmonised practices set out in the standards, but who do not concern themselves with the potential harmfulness of the resulting system. This disclosure obligation is similar to what is required for AI used for recruitment in the state of New York in the US, with regard to algorithmic biases. However, studies have shown that although providers disclose the results of their bias audit, as required by law, these systems still contain many biases. Providers do not mitigate these biases because they are not obliged to do so, only to disclose them (Groves et al., 2024). Simply focusing the standards on disclosure and good management practices could have a similar effect if they do not require anything about the outcome of these assessments.

The question of what standards should contain, both in terms of their technical level and the way in which they address normative issues, is a political choice. The ecosystem of standardisation seem to tend to prefer governance standards that leave room to company to decide what process they want to implement. But as Yates and Murphy (2019) notes, the shift to quality management standards has led the world of standardisation to become *"ironically less focused on the social good"*. Indeed, large companies appreciates these open-ended standards because they often do not even need to change their processes to meet them, they simply have to generate documentation (Choi, 2024). If too much room is left to companies, without ensuring that the choices they make indeed guarantee a high level of protection to fundamental rights, the safeguards of standards and subsequently of the AI Act will fail.

9.3.4 *Getting out of the neutrality discourse*

The standardisation sector presents itself as an area of technical expertise, producing guidelines and requirements that are supposed to be based on science. Standardisation experts generally insist that technical considerations alone determine standards (Loya and Boli, 1999). Their legitimacy to define the best practices to follow is based on the fact that they are professionals who have knowledge of the field and can be considered "technical experts". But this reliance on expertise poses problems for three reasons. First, it fails to recognise that normative choices are embedded in science and technique. Second, we have seen that this expertise, at least in the case of AI, lies not so much in science itself, but rather in the processes of SDOs. Finally, these organisations prone governance principles that are erected as universal, but which itself does not properly respect.

In the words of Loya and Boli (1999), the authority of standards bodies *"derives from the truth presumably embodied in science and technique, from the righteousness presumably embodied in the principles of governance by which the standards bodies operate (equality, fairness, nonpartisanship), and from the presumed self-interest of the lower-level actors that comprise them."* Standardisation is based on the same ideal of neutrality in science that scientists have long endorsed (Harding, 1992). This use of the neutrality argument is not new, as science has often be used in history to justify lobbying activity (Saltelli et al., 2022). Similarly, it is not unusual for standards setters to share the view that technology is non prescriptive. The discourse of the standardisation experts we interviewed bear witness to this, as there seems to be a mistrust of normative questions[26]. Many experts fail to recognise that what they choose to include in standards is in itself value-laden. For Winfield (2019b), there is no standard which is not an ethical standard. Failure to recognise that no technical standard is neutral runs the risk of experts not being sufficiently critical of the choices they make.

But this neutrality discourse is also present in the marketing of standards, where the expertise of the people who develop standards is often put forward[27]. However, we have seen that many of the experts who develop AI standards are not computer

[26] See Chapter 8, Section 8.5.2.

[27] See Chapter 4, Section 4.2.1.

science experts, but governance experts who are familiar with standardisation processes[28]. In fact, some of them are working on standards in several fields at the same time – AI but also quantum computing, software, etc. There is a disconnect between what is displayed by SDOs as a selling point for standards and the reality of profiles in standards development. For standards outside of AI, one could argue that what matters is not technical expertise but field expertise. However, as far as AI is concerned, the discussions in standards setting are captured by cross-sector BigTech companies. This is coherent with the objective of AI standards today: to be horizontal and high level in order to be applicable to any AI technology. This pushes aside sector-specific considerations and places less value on field expertise, which depends on the context of deployment of AI technologies.

According to Loya and Boli (1999), the standardisation sector *"displays principles of universalism, rationality, and homogenisation to an extreme degree"*. This is best illustrated by the World Trade Organisation (WTO) principles for international standardisation processes (WTO, 2000), to which all the SDOs presented in this thesis adhere. These principles include for instance "transparency", "openness" and "impartiality", to which the same criticisms can be levelled as to the ethical principles present in AI ethics charters: their generality and imprecision make it very easy to adhere to them, but more difficult to agree on how to respect them in practice. SDOs have even been accused of failing to respect some of these principles, such as the principle of transparency, as the inner workings of these entities and their processes are still highly opaque[29]. Vague principles and statements can also be found in the CEN-CENELEC code of conduct (CEN-CENELEC, 2018) that members and observers participating in the work of CEN-CENELEC must respect. These include *"work for the net benefit of the European community"*, *"behave ethically"* and *"respect others"*. Incorporating these principles into a code of conduct or international agreement enables the SDOs to protect themselves against criticism, for example by citing these documents when accused of opacity or partiality. Furthermore, the notion of consensus, which is enshrined in these two texts and is at the basis of the operation of many SDOs today, can also be used as an excuse when the content of the standards is criticised for being partial. Experts and organisations have argued that if an element is included in a standard, it is because it has been the subject of a consensus, which justifies its adoption and legitimacy. However, what is not as visible is that this consensus is the work of a very small number of people[30], with a majority of big industrial stakeholders[31].

The neutrality discourse held by the SDOs and their experts can be harmful, as it conceals certain problems of standardisation, such as the normative choices made by experts, the influence of economic and political incentives on standards, the lack of sector-specific technical contributions and the overall dysfunctions of an opaque and partial system. Whether the content of standards is more technical or more focused on governance, it is important to recognise that the content of standards is biased in favour of the interests of experts, in order to understand how standardisation could be improved.

[28] See Chapter 8, Section 8.4.1.

[29] See Chapter 7, Section 7.8.1.

[30] See Chapter 8, Section 8.6.2.

[31] See Chapter 8, Section 8.6.4.

9.4 AI standards could mark the beginning of a fourth wave of standardisation

9.4.1 AI standardisation has elements from the third wave of standardisation

The history of the evolution of standardisation through the ages is made up of numerous waves and trends that have shaped the system we know today[32]. In particular, what Yates and Murphy (2019) has called the "third wave" of standardisation, which dates from the 1980s and continues to this day, relies on four key elements[33]: (i) a diversification of standards setting organisations, with groups of professional organisations taking over the role of traditional organisations; (ii) a diversification of profiles within these organisations, with more representatives from the civil society; (iii) a diversification of types of standards, with more "quality management" standards and less "product requirements" standards; and (iv) a diversification of processes of standards setting, with a decision making process that resembles deliberative democracy.

At first glance, AI standardisation meets all the conditions to be considered as a type of third wave standardisation. Firstly, although traditional standards bodies such as ISO still have a strong presence in AI standardisation, particularly with the extensive work of SC 42, we are indeed seeing a number of other organisations setting standards for AI, particularly professional organisations such as IEEE. Secondly, the example of CEN-CENELEC shows that, although civil society groups are not yet as present as the private sector in AI standardisation, they are more widely included insofar as AI has a strong impact on the fundamental rights of individuals and standards grow increasingly closer to normative issues. Thirdly, the production of recent AI standards, such as ISO/IEC 42001, and numerous initiatives within CEN-CENELEC JTC 21, show that the trend towards quality management standards – or more generally "governance" standards – is far from over. Like ISO 9001 in its day, 42001 could become the world's most widely adopted standard for standardising AI production in companies. But this trend is not limited to ISO, as the National Institute of Standards and Technology (NIST) in the US is also tending to focus more on open frameworks, such as the NIST Risk Management Framework (RMF) for AI (Choi, 2024). Finally, we note with AI standards that the standardisation process based on consensus is predominant at both ISO/IEC and CEN-CENELEC, and that the discourse these organisations hold is a unifying one, eager to display a high degree of legitimacy. Yet, they remain highly political and diplomatic spaces, rather than pure technical ones.

These four tendencies are not independent from one another, as we witness a diversification of standard setters in general, whether in terms of organisation types or individual participation. Similarly, the diversification of profiles, including towards less technical profiles, has accelerated the proliferation of standards that are also less technical, and has forced processes to become, at least in appearance, more democratic, in order to listen to this plurality of voices.

9.4.2 AI standardisation goes further than the third wave

However, while AI standardisation appears to tick the boxes of the third wave of standardisation, it also has some distinctive elements which lead us to believe that a fourth wave of standardisation could be on the horizon.

[32] We have summarised this history in Chapter 4, Section 4.4.1.

[33] Yates and Murphy (2019) mostly mentions point (i) and (iii). The two other points are very complementary but should, in our opinion, be separated from the former two.

To begin with, despite the very dominant position of international standards organisations such as ISO and transnational professional organisations such as IEEE, national and regional standards initiatives are also gaining in visibility. This is the case with CEN-CENELEC, which is seeking to break free from the influence of international standards and bring a more European touch to standards by developing standards more respectful of "European values". In addition, national institutes, such as the NIST in the US, are developing their AI frameworks independently of ISO and other international bodies, include them in local policies and laws[34], while gaining in visibility outside their home country. But this trend towards local standards is unlikely to replace international standards as these two types of initiatives serve different objectives. We therefore believe that a return to the establishment of national or regional standards, as in the first wave of standardisation, is highly unlikely. However, we could see a proliferation of AI standards, with many different entities producing their own framework. This was already the case with the AI ethics charters which, in addition to company charters, have also been drawn up by various government institutions in different countries or regions, such as the HLEG in Europe, and by supranational organisations such as the OECD[35] or UNESCO[36]. Like AI ethics charters, this trend where each entity develops its own AI standards could start with companies. We are already seeing the emergence of private AI standards from a single company, such as Microsoft's (Microsoft, 2022). If this trend towards the diversification of sources continues with future AI standards, we will have the first element to consider it as a new wave of standardisation.

If non-technical profiles such as civil society groups are more included with AI, we also see a diversification of the "technical" profiles, where company experts are less experts in AI itself and more experts in the governance of standardisation processes. This is mainly due, at least in the context of AI, to the inclusion of experts from BigTech companies, who aim to disseminate corporate objectives in the development of standards. To this end, they take on various roles within these standards bodies, getting involved in initiatives related to AI, but also in a variety of other areas. This diversification of profiles towards "governance" experts is the second elements of this fourth wave of standardisation.

Additionally to the proliferation of quality management standards, we are seeing the emergence of new areas of interest, focusing more on normative issues, such as standards for ethics or fundamental rights. If the first examples of standards on normative issues goes back to the General Data Protection Regulation (GDPR) and its standards on privacy and data management, and to the IEEE standards on the ethics of robotics (Bryson and Winfield, 2017), it takes on a new dimension with AI, particularly in Europe where CEN-CENELEC is actively working on these subjects. But Europe is not alone in working on standards relating to normative issues. IEEE, for instance, is continuing to publish new standards on the ethics of autonomous systems, and ISO/IEC is also tackling the social issues associated with AI. More specific standards linked to ethical principles combined with algorithmic practices are also being developed, such as fairness standards. This new normative dimension is distinctive of AI standards. Although technical choices have always been value-laden, the topics that AI standards choose to address are more political than before and raise questions about their practical implementation.

Finally, while the processes for developing AI standards are based on consensus[37] and all interest groups in society are called upon to participate, this apparent democracy is hindered by the stronger participation of large companies. While

[34] NIST standards are cited in official communication of the US Department of State (Bureau of Cyberspace and Digital Policy, 2024).

[35] Organisation for Economic Cooperation and Development.

[36] United Nations Educational, Scientific and Cultural Organization.

[37] At least in traditional SDOs such as ISO/IEC or CEN-CENELEC.

this was already known in the literature covering the third wave of standardisation (Mattli and Buthe, 2003; Brunsson and Jacobsson, 2002; Büthe and Mattli, 2011), with AI, traditional sector-specific corporations are being replaced by BigTech companies, with horizontal views on AI, which are deliberately keeping discussions at a higher level[38]. In addition, these players have understood the power of this structure based on national bodies and have infiltrated these different national groups to ensure their dominant position overall in the SDOs. This crushing of other stakeholders seems to be characteristic of AI standards, or at least more prevalent than in other areas, due to the gigantic economic power held in particular by the BigTech, as a mirror of their position in our modern society. To fight against this stranglehold, in Europe, the Commission could decide to set aside the ESOs and to develop its own standards: the "common specifications" for AI (Art. 41 AI Act). These specifications will probably not be developed using traditional consensus-based processes, but rather the Commission could choose who develop them, thus hijacking these deliberative processes[39]. With AI standards could therefore come the end of consensus-based standardisation processes, for which we saw the limits in the current power dynamics within SDOs.

9.5 *The obstacles faced by the European standardisation system force the Commission to intervene*

AI standards inherit a system that was not designed to support legislation or to deal with a technology with so many social implications. As a result, it may be ill-suited to deal with AI-specific issues. But in this situation, the Commission still has several levers for action.

9.5.1 *A lack of trust in standardisation*

The European standardisation system faces a number of obstacles today, such as the length of its processes, the business model of selling standards even when they are harmonised, the adoption of international standards to support European legislation and the stranglehold of large companies on the content of standards. Senden (2017) therefore questions the *"constitutional fit"* of the European standardisation system, as there do not appear to be sufficient safeguards with regard to the delegation of power, competence and implementation. There is therefore a lack of trust in standardisation, including from the European Commission.

A first indicator of this lack of trust in the standardisation system is the gradual integration of common specifications into EU regulations, including the AI Act[40]. These common specifications come alongside a long list of various means of compliance created by the AI Act, with different levels of recognition: codes of practices, codes of conducts, delegated acts, implementing acts, and guidelines[41]. The multiplication of these frameworks shows European Commission's desire to have more control on the technical content of the AI Act. With the exception of codes of conduct and guidelines, all documents will be superseded by hENs, but only if hENs exist and meet the Commission's expectations. The variety of documents planned therefore reveals the need for safety nets in case hENs fail.

The lack of trust in standardisation is also reflected in the sidelining of ETSI from the AI Act's standardisation request. [P2] recounts: *"[ETSI's] governance is not very well accepted by the European community. They are extremely efficient, but at the same*

[38] See Chapter 8, Section 8.6.4.

[39] For more information on common specifications, see Chapter 7, Section 7.9.4.

[40] For more information about the integration of common specifications into EU law, see Chapter 7, Section 7.9.4.

[41] For more information on these frameworks, see Chapter 3, Section 3.5.5.

*time, in ETSI, whoever pays more has more power"**. The Commission therefore fears that CEN-CENELEC will become like ETSI: very opened to international voices and controlled by the stakeholders who have the most money to invest.

Although CEN-CENELEC is far from being in the same situation as ETSI, its relationship with the Commission is also shifting, with EU officials becoming increasingly involved in standardisation and exercising stricter control over hENs in the standards development phase, whereas previously they were only involved in validating the end of the harmonisation process with the publication in the OJEU. The Commission notably tries to communicate as best it can with the ESOs on the types of deliverables it wants: standards that are adapted to EU legislation and sufficiently precise (Soler Garrido et al., 2024), while being developed by a diverse range of stakeholders, including academics, SMEs and civil society organisations[42].

Even after the standards have been developed, the Commission could still strengthen its control over the harmonisation process. Indeed, some experts believe that the Commission can only base its assessment on a strict comparison between the scope of hENs and the scope of the corresponding EU regulation and standardisation request (Ebers, 2022). Yet, the ecosystem newly created by the AI Act, including the AI Office and its 140 employees – among which are technology specialists – could help bolster the Commission's manpower for reviewing standards.

9.5.2 *A mismatch between EU demands and what is possible*

AI standards are expected to be both technically precise, but also address all AI systems at the same time, while encompassing normative issues such as the protection of fundamental rights, and complementing the law. They are the cornerstone of the European strategy for AI regulation, but are unable to meet expectations because of how much they are required to do and the complexity of the task at hand. This raises the question of whether we are asking too much of standards. There is a mismatch between the possible outcome of standardisation, i.e. what the standardisation system can do, and what is expected by Europe.

One example of this mismatch is the promotion by EU institutions of a more diverse standardisation system, with more academics and civil society representatives[43], while initially standardisation is understood as a consensus between private actors. This demand goes hand in hand with the desire to integrate ethical, social and fundamental rights issues into standardisation and legitimise the ESOs.

But this diversification is not always well received by standardisation experts. A study from the United Nations notes that there is in standardisation a *"prevalent perception that including human rights considerations would hinder efficient, speedy standard development and implementation processes, as it would require building new expertise and the participation of more actors"* (OHCHR, 2023). We have also witnessed this discourse, with experts, even in humanities, believing that civil society stakeholders *"follow an agenda that is highly non-technical"* and which *"slows down the process [of standardisation]"* [P16].

Yet in the context of AI standards as requested by the Commission, i.e. dealing with fundamental rights issues, the involvement of these stakeholders is essential. In addition, the participation of academics allows standardisation groups to have more experts in computer science without depending on industry to decide what should be considered the state of the art. We believe that all these perspectives are needed to develop standards that reflect not just what the industry *wants* to do, but also what it *should* do.

[42] See Chapter 7, Section 7.9.1.

[43] For more information, see Chapter 7, Section 7.9.1.

9.5.3 *A variety of possible improvements*

Rethinking the system

This mismatch between the EU's needs and the solutions that the ESOs can provide raises the question of alternatives to this standardisation system. We could imagine a system where the Commission recreates its own standardisation bodies. In the past, the Commission has considered the idea of creating an independent "European Agency for Standards" to manage the standardisation process in place of the ESOs (European Commission, 2011b). This would enable the Commission to have a closer relationship with the standards body, avoid the rejection of standardisation requests and have standards closer to the Commission's concerns. However, the Commission acknowledged that this could only be done with the support of the NSBs and that it would not necessarily reduce development times. In the end, the idea was not pursued for cost reasons, with the Commission stating that this solution *"would certainly lead to substantial additional costs for the EU-budget"*[44]. Indeed, in addition to the structural upheaval it would create, it will probably not be sufficient to make this new body legitimate. A simpler solution would be to improve the current system brick by brick.

On the business model of standards

But even if the current standardisation system is experiencing some difficulties, at least as far as AI standards are concerned, there is still room for improvements. First of all, the business model for standards has recently been called into questions by case laws such as the Public.Resource.Org case[45]. Although the decision from the Court of Justice of the European Union (CJEU) does not entail the systematic publication of hENs free of charge, it may set a precedent for more open publications. However, the issue of the copyright on standards has yet to be addressed, as it may conflict with the right of access to legal acts such as hENs. The business model of hENs needs to be further clarified, particularly with regard to whether or not it is suitable to support European law. In addition, the SDOs are currently thinking about improving their products and services, with "smart" standards that could be implemented directly by companies[46]. The emergence of this new type of standard could revolutionise the standards business model and may solve the issue of right of access. Nevertheless, we recommend improving access to these documents as they have strong legislative power and can have a direct impact on citizens' rights, especially when standards actively attempt to address normative issues such as fundamental rights.

There also appears to be a funding problem, with the Commission stating in the early 2010s that it contributes to almost half of the total income of the three ESOs, even though hENs account for a small portion of their deliverables (European Commission, 2011b)[47]. The same applies to AI, with several grants helping to fund AI standardisation efforts[48]. The Commission could decide to make this funding conditional on free access to standards. However, such conditions would mean that the Commission and the ESOs recognise that their relationship is similar to outsourcing, with the standardisation request being similar to a service contract.

[44] The additional cost is estimated to be at least €20,000,000.

[45] For more information, see Chapter 7, Section 7.3.

[46] According to Baeva et al. (2023), smart standards are machine readable standards that could be automatically evaluated and verified.

[47] 11.9% of CEN-CENELEC standards are cited in the OJEU today (CEN-CENELEC, 2024). It should be noted, however, that these two studies were carried out more than ten years apart, but we were unable to find more recent figures.

[48] Such as two "Single Market Programme" grants (European Commission, 2023d,c)

On participation in standardisation

While the European Commission considers that *"meaningful participation"* (European Commission, 2011b) in standardisation requires a baseline of approximately 20% of a person's time, this is currently much higher in AI standardisation, as there are very few people actively involved compared to the task in hand. Indeed, while standards are necessary in our modern society, standardisation is an activity invisible to most people (Spivak and Brenner, 1993), citizens, scholars and small companies alike. Solutions to this lack of interest include launching awareness campaigns, teaching the role of standardisation at school – in particular in engineering and law programmes, providing additional funding to enable experts to participate, and so on. To improve effective participation and maintain people's interest in standardisation once they are registered, training could be provided, for example by NSBs.

A key stake is the participation of European SMEs in the ESOs. SMEs find it difficult to benefit from and become involved in standardisation, as they come up against a number of obstacles, such as a lack of awareness of the existence and importance of standards, access to or understanding of standards, and the cost of investment (de Vries et al., 2009). Several solutions were proposed by the standardisation ecosystem and scholars for the better inclusion of SMEs. (de Vries et al., 2009) proposes the development of national strategies for the education on standardisation, the reinforcement of communication efforts specific to these stakeholders, or the possibility to reduce costs to buy standards. More recently, for AI standardisation, the working group 1 of CEN-CENELEC JTC 21 created a Task Group dedicated to "inclusiveness", in charge of raising awareness about JTC 21 activities and bring new stakeholders to the discussions (CEN-CENELEC, n.d.a).

Since standards are to tackle technologies that have a high impact on citizens rights, the involvement of civil society stakeholders is also key to the preservation of these rights. However, despite the Commission's calls for more diversity, the participation of these stakeholders remains sparse. Ebers (2022) proposes that societal stakeholders should be granted voting rights in the development of international standards, rights of appeal, and unlimited access to technical bodies and advisory groups, as well as to existing standards, without charge. Galvagna (2023) proposes to amend Regulation 1025/2012 on standardisation (European Parliament and Council, 2012) to include more participants in Annex III who could be funded by the Commission and participate in the work at the ESOs, to fund more individuals from civil society organisations with the StandICT system[49], and to create or fund a central hub to support civil society participation.

Finally, to include more academics, solutions include greater funding, including from StandICT, as well as recognition for individual work on standards. Universities and research centres could also encourage their researchers to take part in standardisation by offering them specific job opportunities and put in place communication campaigns to highlight the importance of their work. In general, a better understanding of standardisation by the academic system could improve the recognition and valorisation of standardisation activity.

[49] For more information on StandICT, see Chapter 4, Section 4.6.5.

On transparency

The legitimacy problems encountered by ESOs are mainly due to their opacity and lack of transparency, both internally and externally. Indeed, the ESOs do not usually communicate on who their experts are or what they are working on (Perarnaud, 2023). People working in the field of standardisation also point out that it is sometimes difficult to know which company a certain expert works for, as it is not compulsory to disclose this information and, on a voluntary basis, it is not always easy to get people to comply and many refuse [P5]. Indeed, experts are supposed to represent their own opinion, not that of their company or entity of origin. In practice, however, it is difficult to distinguish between the two[50].

Measures could be taken to make the standards development process more transparent. If minutes are already taken, they sometimes do not contain all the underlying information and debates. We therefore suggest that the meetings be recorded and that the recordings be made available to participants in the standardisation process. Affiliations should be made known at standardisation meetings to enable experts to understand the position of their colleagues and to identify lobbying activities. As far as external communication is concerned, we have already suggested that the names of the main contributors should appear on the standards, particularly in the case of academics. If this solution is rejected, an alternative might be to reveal only the names of companies taking part in standardisation work, not necessarily the number of experts they have, but at least the fact that they are involved and on which projects. However, this can prove challenging as standardisation experts come and go from every project all the time and it is hard to quantify the influence they had on a given work. A last solution could be to simply display the names and affiliations, or alternatively the affiliations without the names, of all the members of a given organisation or working group.

All these solutions could help to improve the current system without revolutionising its structure. They do not require major investment and should not be too complicated to implement.

9.5.4 Deadline for standards push the ESOs to consider international standards

Many of the experts we interviewed criticised the lengthy processes of standardisation which hamper the drafting of standards[51]. Today, stakeholders admit that drafting standards, either harmonised or not, take about five years [P5]. But that does not mean that after five years, the standardisation working groups will necessarily produce a standard. Sometimes, processes get blocked.

One of the solution put forward by the Commission in the early 2010s to reduce development time, was to simply put a strict deadline for the ESOs to draft hENs in the hope that processes will be accelerated. In return, the Commission committed to deliver standardisation requests to the ESOs more quickly (European Commission, 2011b). However, today, we see that this strategy has proven inefficient, at least for AI standards. The Commission delivered the request before the final publication of the AI Act, which enabled CEN-CENELEC to start working on standards early on. At the same time, the request provided for a deadline in early 2025, only three years after the request was accepted by the ESOs. However, the experts point out that having to work with both a legal text under construction and a standardisation request dating from the first proposal caused some conflicts, as it was not clear

[50] See Chapter 8, Section 8.6.4.

[51] See Chapter 8.

which version of the text or request the ESOs should refer to [P5]. Furthermore, the deadline imposed by the Commission did not encourage the ESOs to improve their internal processes, but only gave stakeholders who were opposed to the development of new standards more arguments to push for the adoption of international standards.

The adoption of international standards instead of the drafting of new European standards in the context of AI was envisaged early on. Before the standardisation request and the forming of an official JTC on AI, CEN-CENELEC experts on AI were grouped in a Focus group. In a response to the White paper on AI, the focus group explains the role that standards can play in an upcoming European regulation (CEN-CENELEC, 2020). They notably advocated for the adoption of ISO/IEC foundational standards as a reference for upcoming European standards. They also advocate for the consideration of ISO risk management standards for the risk-based approach. This desire to rely on international standards was also displayed in the Focus group Roadmap (CEN-CENELEC, 2020). As confirmed by the experts we interviewed, some stakeholders, such as BigTech companies, supported the creation of a European JTC that would be an *"empty shell"**, simply adopting existing international standards without carrying out any additional work [P5]. These arguments were backed up by the fact that international standards were already published and that, if European work was to start from scratch, it would take a lot of man power, lead to overlaps and inconsistencies with international standards and standards might not even be delivered in time for the implementation of the AI Act. The adoption of international standards as hENs is, however, not new. Indeed, 28% of all hEN from CEN are adapted from ISO standards, and 69% for CENELEC and IEC (CEN-CENELEC, 2024). For instance, the medical device regulation, often cited by EU officials as the most convincing example of the NLF's success, has almost all its harmonised standards derived from ISO[52]. With the support of the European Commission, experts opposed to BigTech managed to get the JTC to reach a consensus on the inadequacy of international standards alone to cover the AI Act. However, there are still major disagreements today that slow down standardisation work, including on the adoption of international standards for risk management.

The example of the AI Act therefore shows that, today, the question of the timing between standards and European legislation is not entirely resolved. However, we do not believe that the Commission could have done things differently. Standards have been rushed through, as has the AI Act, as technology evolves rapidly and regulation must strive to keep up. Ideally, for future regulation, the Commission should try to give more time to standardisation experts, or give them the means, in terms of staff and money, to develop standards more calmly.

9.5.5 *If standards are not ready in time, the Commission will have to make a strong political choice*

The AI Act places so much emphasis on technical standards that it is possible to take them for granted and forget that they do not yet exist (Pouget, 2023). The Commission's deadline of April 2025 is considered at best ambitious, at worst unrealistic (Perarnaud, 2023). Several voices explain that it is still unsure when the hENs will be ready, although it will be, in any case, too late for the initial deadline[53]. Experts now tend to think that a postponement to 2026 is inevitable. Nevertheless, standards must be ready several months before the provisions relating

[52] A list of these standards can be found at: (European Commission, n.d.l).

[53] See Chapter 8, Section 8.5.1.

to high-risk AI systems come into force in August 2026, to give companies time to comply (Soler Garrido et al., 2024). However, this obvious difficulty on the part of CEN-CENELEC in delivering standards on time is at odds with the reassuring tone of the European Commission. In 2021, the Commission announced *"a large set of relevant harmonised standards could be available within 3-4 years"* (European Commission, 2021b) and, in a recent webinar, EU officials remained positive about the ability of the ESOs to meet the deadlines (European Commission, 2024a).

But if the standards are indeed not ready by April 2025, or even 2026, which now seems likely, the Commission will have to decide what to do. A first option could be to publish parts of the hENs in the OJEU, showing what the ESOs have achieved by the deadline. This would maintain the *status quo*, but runs the risk of creating loopholes in standards that providers could exploit given the legal power of hENs. Further work would help fill these gaps, but they would remain for a certain period of time. A second solution would be to wait for a complete version of all hENs. However, this could take years as some items, such as cybersecurity requirements, are far from ready [P16]. Meanwhile, alternative frameworks, which the Commission does not control and which may come from the industry or any other entity worldwide, could emerge and be used by companies to assess their conformity. Furthermore, there is no guarantee that the standards which are being developed will satisfy the Commission and pass the HAS assessment. hENs may therefore never be published in the OJEU and never be granted legal power. This situation, in which a hENs has been offered for citation in the OJEU but was refused, has already occurred in other fields, but remains fairly rare[54]. Finally, the Commission could choose to develop common specifications to fill the gaps during the wait for hENs. However, it is not clear exactly where the Commission would find the manpower and how quickly this could be achieved. The legitimacy of common specifications, which would be drafted *"on the back of an envelope"** [P9], hijacking the consensus-based standardisation processes, could also be compromised.

There are two questions that the Commission must consider. First, what will be the impact on the implementation of the AI Act? Indeed, further economic and legal studies will have to confirm whether the publication of parts of standards is better than nothing, or whether the legal power of hENs is too important to be left to semi-finished work. Second, what will be the impact on the EU's legislative framework and its legitimacy? Indeed, by waiting past the deadline, by not publishing hENs at all, or by choosing to draft its own specifications, the Commission would be acknowledging the failure of hENs and of the European standardisation system as a whole. At a time when all eyes are on the Commission to monitor the implementation of the newly adopted legal text, such bad press could have consequences far beyond AI, as it could be seen as a sign that the current standardisation system is no longer fit to support EU legislation. The decision that the Commission will have to take is therefore highly political and will reveal the extent to which it still believes in the current system.

9.6 Fixing AI standards by refocusing on vertical sectors

9.6.1 European AI standards are horizontal because they follow the AI Act

When we spoke with standardisation experts, they explained that AI standards are very high-level today and rarely go into technical details, such as mathematical

[54] 6.5% of CEN-CENELEC standards which were offered for citation are currently not cited, either because the process is underway or because it was refused (CEN-CENELEC, 2024).

definitions, metrics or thresholds. They rather remain fairly general, with definitions in natural language. We felt that academics and computer scientist experts in particular were frustrated by the lack of technical content. The experts identify several reasons for this general approach to standards, depending on the type of stakeholder to which they belong. Industry experts usually argue that the methods and metrics are not sufficiently mature, and that including them in a standard would be taking a risk: the standard might not work in practice, for example, or might quickly be rendered obsolete by a method that was recently discovered. In general, these industry experts are more focused on implementation. On the contrary, academics tend to believe that industry experts do not want standards to be too technically precise, because less stringent requirements would allow them to be less constrained by regulation. These academic experts are therefore more concerned with ensuring that standards are close to the technical state of the art in research.

However, there are also other reasons why standards remain very general. First of all, AI standards are still in their infancy, with the first to be developed being ISO/IEC standards from 2018 onwards. As standardisation work has to start somewhere, it is only natural to have broader standards at the beginning and more specific requirements later on.

Second of all, the term "AI" is itself quite broad and covers a wide range of contexts, both in terms of (i) technology, and (ii) sector of application. Indeed, AI can be a simple machine learning model such as a classifier, a rule-based model, a deep neural network used for Natural Language Processing (NLP) or another network with a different structure used for image recognition, a reinforcement learning model, and so on. The same image recognition technology, for example, could be deployed in healthcare to detect cancer, in an autonomous vehicle to analyse road signs, or simply to carry out an internet search. These different contexts imply different ethical stakes and the requirements for AI systems are likely to vary depending on the technology and the sector of use.

However, despite these contextual differences, the AI Act itself is very horizontal. Although it lays down different requirements for AI systems depending on their risk category, the categorisation of systems is not strictly based on technology or sectors. Systems can always fall into a certain category depending on their context. AI systems used for education or justice, for example, are immediately classified as high-risk because these sectors are included in Annex III. Systems used in the healthcare sector will also be classified as high-risk if they fall under the medical devices regulation listed in Annex II. Similarly, some technologies fall directly into one risk category. For instance, biometric systems are automatically classified as high risk and, in certain circumstances, can also be classified as unacceptable risk. However, although risk categories are sometimes constructed by listing different technologies or sectors, the requirements applicable to them are neither technology-specific nor sector-specific. All high-risk AI systems, whether they are used for education or healthcare, whether they use NLP or image recognition, will have to comply with the same requirements in terms of data quality, transparency, accuracy and so on. This horizontality is specific to the AI Act, as other EU safety regulations are product-specific (Ebers, 2022). One of the experts we interviewed, [P16], sums up the situation as follows: *"the major challenge we are facing [in standardisation] is that the AI Act is a product safety regulation approach. And AI is not a product. AI is a component of products"*.

Policy documents prior to the AI Act show that the European Commission's approach with the AI Act was first and foremost to avoid the duplication of sector-specific legislation that would each impose their own requirements in an inconsistent manner[55]. To remain in line with the AI Act's approach of having the same requirements for all high-risk AI systems, regardless of their technology or sector of use, the standards that will be harmonised and support the legislation must also be horizontal. They must therefore address in a general manner each of the requirements of the AI Act – transparency, accuracy, etc. This horizontality is even cited by the European Commission as a desirable characteristic of AI standards (Soler Garrido et al., 2024). If CEN-CENELEC had proposed a standard by technology or by use case, this would probably not have passed the check of the HAS consultant, who assesses in particular whether the standards being harmonised are in line with their respective legislation.

9.6.2 *Standards need context*

In addition to its horizontal nature, the AI Act draws heavily on ethical charters and policy documents that are based on the concept of "trustworthiness", i.e. a list of different ethical and technical properties required of AI systems. However, these documents are usually quite broad, simply defining principles or ethical-technical properties in general. On the contrary, a part of these documents focuses on very technical elements, sometimes by limiting the field of study to one technology –such as facial recognition, or one sector – such as education. The conclusion seems to be the same for ethics charters as for standards: it is possible either to stay at the level of AI in general and define broad properties without going too much into technical detail, or to dive deeper into a use case and define more technical elements, give more precise definitions, metrics and methods more aligned with the context of use.

For instance, the performance of an AI system is calculated very differently depending on whether it is a classifier or a language model. In the case of a classifier, a simple measure of accuracy or f1-score might suffice. For language models, many metrics might exist depending on which task the model is trying to perform. Indeed, models can focus on processing tasks such as sentiment analysis, text classification, inference or semantic understanding. They could also generate, translate or summarise text. Each of these tasks will therefore require a specific metric (Chang et al., 2024). Similarly, fairness metrics are different in regular classification models, and in facial recognition models for instance. While classification tasks usually use group fairness measures such as demographic parity or equalised odds, fairness metrics in facial recognition are often based on a comparison between the error rates of the two groups on which the system performs best and worst, or on the average of all error rates (Grother, 2022).

In some cases, some trustworthiness criteria might not even be relevant. Fairness, for instance, is more important in recruitment than in industrial production (Baeva et al., 2023). In recruitment, ensuring fairness involves ensuring that performance is similar between groups of people, for example between men and women. In this case, fairness and good practice in mitigating bias will have a direct impact on fundamental rights. On the other hand, group fairness in industrial production only concerns the correct functioning of the algorithm for different types of construction parts. It will therefore have no direct impact on individuals.

[55] See Chapter 3, Section 3.2.4.

9.6.3 *Other harmonised standards are vertical*

The AI Act is one of the first product safety regulations adopted under the NLF not to be strictly limited to a single sector. For comparison, other regulations, such as for toys or medical devices, refer to a particular industry: the toy industry or the healthcare industry. The standards referenced in the OJEU are therefore also vertical in these sectors. This verticality of previous standards is not a coincidence: according to Choi (2024), standards work best for narrow tasks for which they can give quantitative attributes. This is particularly true of software standards, which have become less effective since they have being reduced to large open-ended frameworks. Choi (2024) therefore calls for standards to be refocused on more specific tasks in order to make them more effective, and for very broad quality management standards that leave too much flexibility to industry to be banned.

One of the only exception to this verticality of NLF regulations is Regulation 2024/1781 on ecodesign requirements for sustainable products (European Parliament and Council, 2024c) – formerly Directive 2009/125/EC (European Parliament and Council, 2009a). The 2024 Regulation is too recent to have hENs, and there are no references to hENs published in the OJEU under the 2009 Directive. Indeed, the former Ecodesign Directive, as well as the new Ecodesign Regulation, are what European institutions call *"framework legislation"*. In the words of the European Commission (n.d.e), this means that *"concrete product rules will be decided progressively over time, on a product-by-product basis, or horizontally, on the basis of groups of products with similar characteristics"*. These concrete product rules are published in the form of Commission's implementing measures, such as for computers, dishwashers, ovens, fans, washing machines, etc[56]. As a result, although the regulation is horizontal, the hENs for ecodesign requirements are sector-specific on these products. A few standards retain this horizontal approach to enable the alignment between vertical standards, but they are not harmonised. However, this raises the problem of the interplay between horizontal and vertical standards, as well as a lack of clarity surrounding the role of horizontal standards (Bundgaard and Huulgaard, 2023). Horizontal ecodesign standards can serve as a *"shopping list"* and ensure alignment (Bundgaard and Huulgaard, 2023), but discussions are still ongoing to see if they can be used directly for certain products or if product standards should be developed first (Schischke et al., 2022).

9.6.4 *Developing vertical standards will help with various issues*

Refocusing AI standards on vertical sectors could help in several ways. First, it would help solve practical implementation issues by defining technical requirements that are feasible in a certain context if use. It would also support the regulation much more efficiently by declining each legal requirements into verifiable quantifiable criteria, which are impossible to define today at the horizontal level. The result would be more a effective regulation that AI system providers would be obliged to comply with, rather than interpreting in their own way, which could lead to abuse.

Verticality could also be a solution to the integration of values in standards as certain values could be prioritised in a given context[57]. Indeed, the hierarchy between principles is never addressed in ethics charters, nor is the hierarchy between requirements in the AI Act. If in theory, all legal requirements should be respected –

[56] A list of these standards can be found here: (European Commission, n.d.d).

[57] Gonzalez Torres and Ali-Vehmas (2024) give the example of AI for voting which could prioritise the value of democracy.

as well as all ethical principles – in practice, some requirements might be working against each other, such as fairness and accuracy. Contextual technical requirement might help address how a compromise could work for a given technology in a given sector.

The development of vertical standards could also help to solve some of the structural problems facing the European standardisation ecosystem. More vertical and therefore more technical standards could attract more academics and researchers, who are more interested in the technical details of standards. This would also help to involve SMEs, which are generally more interested in the practical aspects of implementing standards in their sector than in the general standards that large companies are currently pushing at European level [P13]. More vertical standards would therefore attract domain specialists, complementing the knowledge of AI experts who may not necessarily have in mind the necessary details for the proper deployment of an AI system in a given sector. Ultimately, this would provide a counterpower to the voice of larger companies, and a greater willingness to publish standards.

9.6.5 *Vertical standards also have their limits*

However, vertical standards are not a miracle solution. First, an approach that would focus on sector-specific standards runs the risk of confusing cross-sector providers who would have to comply to different standards. For instance, a large language model (LLM) used for proof reading, can be used in education, employment, justice, and so on. For each of these use cases, the AI system used will be the same – even if sometimes the model can be fined-tuned to fit the context, and often the provider will also be the same. With sector-specific standards, the same provider would be overwhelmed with different standards for the same product.

An alternative would be to consider "vertical" standards as technology-specific rather than sector-specific. This would notably allow to define terms and metrics that are specific to one area of AI, i.e. NLP, computer vision, expert systems, etc. However, this would still not suffice to define thresholds that also depend on the context of use. Vertical standards would therefore ideally account for both the technology and sector.

Having sector-specific standards also means covering the ten requirements of the standardisation request for each sector, possibly for each technology within that sector. But even then, Annex III of the AI Act which lists the "new" sectors in which AI systems will be considered high-risk, can be amended by the Commission in the future This means that vertical standards will have to be developed for sectors that are not even considered high-risk today. A new technology can also always emerge in one of these sectors, constantly requiring the development of new standards. A vertical approach to AI standardisation is therefore a gigantic task and risks leaving some sectors without standards. The question then becomes what to do in case there is no standard available for a certain technology in a certain sector. Companies could continue to develop their products but would not benefit from any presumption of conformity with the AI Act, which could penalise emerging start-ups and SMEs.

There is also a timing issue: if the development of vertical standards only starts now, they will not be ready when the provisions on high-risk AI system enter into force. This will especially be problematic for providers of biometric systems who, under the AI Act, can either follow harmonised standards or undergo a third party conformity assessment (Art. 43(1)). However, at the time of writing, there is no notified body for biometric systems. This could also pose a problem for Annex I AI systems, for which the provider has the option of opting out of third-party conformity assessment if it applies harmonised standards.

These limits of vertical standards mean that horizontal standards are still needed. However, they would not necessarily provide presumption of conformity, simply cover the gaps in case vertical standards are missing for a specific sector or technology.

9.6.6 Horizontal standards are not necessarily a first step towards vertical ones

As AI standards are still in their infancy, it is difficult to predict what they will look like in the future. Even if standardisation efforts are horizontal today, some others vertical ones may develop in the future. Some experts therefore remain convinced that vertical standards will come after the publication of horizontal standards [P15]. However, defining vertical standards after horizontal standards still raises practical difficulties.

Firstly, having given general elements and definitions could prevent vertical standards from defining their own terms and requirements. As alignment is crucial in standardisation, if horizontal standards define terms – such as accuracy or transparency – in a certain way, it is virtually impossible for other subsequent standards, such as vertical standards, to go against this definition. However, the same term may have different definitions in different contexts. If the definitions are decided at the level of AI in general, they may not work for certain technologies and certain contexts. Secondly, even if vertical standards develop at a later stage, the harmonised standards that will prevail and have legal effect will always be the less precise horizontal standards which have been cited in the OJEU simply because they came first.

Standardisation experts believe that it is too late to change the ESOs standardisation strategy and opt for a vertical approach to AI standardisation [P16]. Indeed, CEN-CENELEC JTC 21 has been working for several years on the horizontal "AI Trustworthiness framework", neglecting vertical ones. A few exceptions are projects on technology-specific standards, such as NLP or computer vision. However, other technologies and sectors have not yet been addressed. While some groups at international level are dealing with systems that would fall under the definition of AI, such as ISO/IEC JTC 1/SC 37 on biometrics, these initiatives are also still in their infancy and European working groups seem to be struggling to progress. Ideally, there would be different working groups developing technology-specific standards, notably for terms and metrics. In addition, there would be one working group per sector impacted by the AI Act, including Annex I on product safety regulation – e.g. an AI working group for medical devices, one for AI in machinery, one for AI in toys, etc. – as well as Annex III on sectors of application – e.g. a working group for AI in education, one for AI in critical infrastructure, etc. In addition to working groups on high-risk AI systems, we could consider working groups on General Purpose Artificial Intelligence (GPAI) models separately in anticipation of a potential standardisation request on this topic[58].

However, such a high number of working groups and standards are simply too high for the ESOs to handle. Even when we discussed the AI Trustworthiness

[58] The Commission is required to issue a standardisation request on GPAI models (Art. 40(2)), but Codes of Practice will already cover these obligations until harmonised standards are available. This means that the timing is less problematic than for high-risk AI systems where there will be nothing until harmonised standards are adopted.

framework with experts, they told us that, at this pace, it will be difficult to pass the HAS assessment so that it could be published in the OJEU [P16]. Not all the items that should be covered will be ready in time, and the framework might not correspond to what the Commission had in mind. We therefore propose that the AI Trustworthiness framework should be adopted by CEN-CENELEC when it is ready, to be a starting point for other standards, but it should not be harmonised. Once this work is finished, JTC 21 should start working on vertical standards that are technology-specific and sector-specific. However, these vertical standards cannot be harmonised since the standardisation request was for horizontal standards on AI and the Commission has insisted on the need for horizontal hENs (Soler Garrido et al., 2024). This solution entails that there will be no hENs published in the OJEU for the AI Act, therefore no standards that gives a presumption of conformity to the law, at least not in the near future, until the Commission decides to issue different standardisation requests for each sector. This lack of hENs would negatively affect especially small companies who do not have the resources to produce their own in-house standards or defend themselves in court when they are attacked for non-compliance. But at the very least, this would avoid the introduction in the OJEU of overly broad standards which would create more loopholes than they would resolve.

9.7 *Europe has made imperfect but commendable efforts to regulate AI*

Europe prides itself with being the first region of the world to adopt a mandatory framework for AI technologies. Despite the efforts made, the text is far from perfect. It will, however, contribute to the emergence of a culture of AI regulation that will only improve in the years to come. In this section, we reflect on the AI Act approach, examining the problems it may raise, the novel approach it helped to initiate and the mark it will leave on AI regulation.

9.7.1 *A contested approach to AI regulation*

Regulating "AI": a political will

Over the last ten years, the EU's general discourse on AI has shifted[59]. At the beginning, AI was only considered a subpart of new technologies that needed to be regulated, including robotics or the Internet of Things (IoT). The shift to AI, particularly with the creation of the High-Level Expert Group (HLEG), was perhaps more political than practical.

AI is a fuzzy family of algorithms, and is hard to define. Some researchers see "AI" as a buzzword used by software providers to sell their products more effectively, engaging in an "AI washing" culture, where anything remotely related to computer programs could be considered AI (Leffrang and Mueller, 2023). Indeed, even just a few years ago, "AI" was not a term that was very popular with computer scientists and researchers, who preferred to use the term "machine learning". The term "AI" was generally reserved for marketing purposes, press headlines, or when applying for funding (Crawford, 2021, Introduction).

The European Commission could have chosen to propose a "Machine Learning Act", where the boundaries are easier to draw, and where the systems primarily targeted by the AI Act – those that pose the most dangers to individuals and society,

[59] For more information, see Chapter 3, Section 3.2.

are at the center of the regulation efforts. On the contrary, it could have chosen to propose an "Algorithm Act", in order to encompass more technologies that could possibly also pose a threat to human rights and which do not currently fall within the scope of the AI Act, such as optimisation methods for scheduling (Abdalkareem et al., 2021)[60]. Yet the Commission has chosen to regulate "AI" specifically, taking a contested word and introducing it into everyday language. This widespread use of the term "AI" is of course not simply the work of the European institutions: the term was already widely used in the first ethics charters of the 2010s, and its popularity has continued to grow ever since, including in scientific publications (see Figure 9.1). But its use by the legislator reinforces the legitimacy of its adoption.

However, even though AI is a more widespread term and technology, there is still a huge amount of work to be done in educating the general public about what AI is and how to use it. The anthropomorphisation of technology does not help to distinguish between what is human and what is not, particularly with the use of terms such as "ethical" or "trustworthy", encouraged by the EU institutions, which are only supposed to be associated with moral beings (Tessier, 2022).

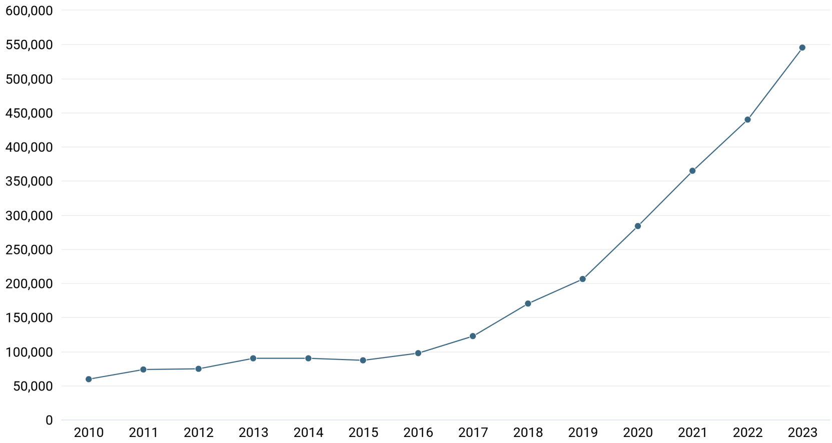


Figure 9.1: Number of scientific publications on AI per year. Source: `https://app.dimensions.ai/`. Criteria of search: "artificial intelligence" in full data.

As the technology evolves rapidly, there is also a risk that the vocabulary and categorisation set out by the AI Act do not stand the test of time. It should be remembered that the GPAI category was added to the AI Act after the release of generative AI models, such as ChatGPT, in late 2022, because the risk classification established by the Commission was heavily based on AI applications and these new systems were impossible to classify in this way, which changed the whole structure of the legal text. The question therefore remains as to whether the AI Act, with a scope of application and a classification system adapted to current AI technologies, will not quickly become obsolete.

Creating trust, embracing the use of AI

The decision to frame AI regulation in terms of the trust we can place in these systems is also questionable. The European Commission explains that the EU's approach to AI regulation will *"give people the confidence to embrace these technologies and encourage businesses to develop them"* (European Commission, n.d.i). The EU's strategy for regulating AI is therefore not to restrict the use of the technology, but to make sure that we use it in a "safe" way. "Trustworthiness" therefore becomes a means of achieving "acceptability" (Laux et al., 2023).

[60] Note that these types of algorithms would have been considered "AI" under the first definition given by the European Commission in its proposal for an AI Act. Indeed, the Commission's definition encompasses *"Statistical approaches, Bayesian estimation, search and optimization methods"* (European Commission, 2021e, Annex I). Some scholars then objected to this broader definition, according to which even simpler algorithms, such as those in a calculator or spreadsheet, could be considered AI (Ruschemeier, 2023).

The acceptability of the technology in the AI Act is, however, strongly linked with the context of application. This distinction is necessary because "AI" is a broad term which encompasses many technologies that do not pose the same problems depending on how they operate and their environment of use. The question of *when* to use AI is perhaps one of the most important for AI regulation. Indeed, some applications of AI do not pose many ethical problems and could even be very beneficial, such as AI for studying climatic phenomena (Rivera Tello et al., 2023), preserving ancient books (Teel, 2024) or protecting endangered species (Müller et al., 2023). However, most applications directed towards the general public do not offer significant benefits that would outweigh the cost of AI to society, democracy and the environment. Rather than simply distinguishing whether a technology is benign or harmful, Narayanan and Kapoor (2024) also suggest looking at which technologies work well and which are *"snake oil"*, i.e when companies are simply riding the AI hype, using it in applications where it cannot work properly, such as for human resources or for predicting health complications. Rather than asking *when*, we should maybe ask *who* should use AI and *for what* purpose.

According to EU institutions, AI can help *"find solutions to many of society's problems"* (European Commission, n.d.i). One of the stated purposes of the AI Act is notably to *"ensure that AI leads to socially and environmentally beneficial outcomes"* (European Parliament and Council, 2024b, rec. 142). The text further mentions *"AI-based solutions to increase accessibility for persons with disabilities, tackle socio-economic inequalities, or meet environmental targets"*. To achieve this goal, Member States are encouraged to *"support and promote research and development of AI solutions"*. This discourse is shared among many actors, particularly those promulgating ethics charters and AI systems "for good". This trend is better known as "technosolutionism", whereby new technologies are presented as the solution rather than the problem (Ferrari, 2020).

But to avoid the many dangers of AI, we will also have to learn *not* to use AI. Many voices have raised these past few years to stop or slow the AI burst. If some of these voices are motivated by economic incentives[61], many are concerned that we are putting our society at risk by continuing this trend. While the AI Act has, at least, banned certain systems, it probably does not go as far as it could (or should), not stopping the AI race but, according to some, encouraging it (du Net, 2024).

An isolated law

Another criticism that could be addressed to the AI Act is its lack of interaction with other European legislative frameworks. Papakonstantinou and De Hert (2024) notes that Europe has made an habit of mimicking the GDPR for the regulation of digital technologies, creating large and independent laws that work in silos and sometimes contradict each others: they call this phenomenon the *"act-ification"* of European regulation.

As European laws on digital technologies are multiplying, it will become crucial to understand their interplay, for instance between the AI Act and data protection laws such as the GDPR (OECD, 2024a), or cybersecurity law such as the Cyber Resilience Act (Burri and Zihlmann, 2023). Indeed, some scholars point to inconsistencies between the AI Act and the GDPR regarding the lawfulness of personal data processing (Bogucki et al., 2022). The AI Act will also have a strong interplay with other EU harmonisation legislation (Annex I) covering "vertical" sectors.

[61] The open letter to *"pause AI experiments"* (Future of Life Institute, 2023) was notably signed by Elon Musk, CEO of Tesla, X (Twitter) and founder of OpenAI, as well as a number of engineers from BigTech companies.

Further clarifications are needed to develop a European legal landscape for digital technologies that remains applicable.

A regulation which is too gentle

The strength of the AI Act to effectively regulate these technologies depends first and foremost on which systems will fall into which category and what will be required of providers. Many scholars have called for these rules to be tightened, whether it be banning more systems or imposing stricter obligations for GPAI models. According to some, the text was subject to extensive lobbying during the negotiation phase, softening it and preventing regulators from imposing stronger controls (Wachter, 2024).

The role of AI providers, who will mostly carry out the conformity assessment themselves, is also strongly criticised by the academic literature (Wachter, 2024). Indeed, the industry will self-regulate, choosing the appropriate safeguards while awaiting possible ex-post control. This is described by Laux et al. (2023) as an "expertocratic model" whereby so-called experts choose what is acceptable.

Now that the AI Act is published, discussions are closed about what can or should be in the law. However, there is still one door open: technical standards, which, if they manage to free themselves from the excessive weight of large companies, could refocus on the sector and specific technology, and impose strict testing methods and result requirements.

9.7.2 *A first step for AI regulation*

Even though some prominent scientists nowadays present AI as a potential threat to humanity (Bengio et al., 2025), according to a report by the World Economic Forum (WEF), AI is not at the top of the list of the biggest global risks (Heading and Cavaciuti-Wishart, 2024)[62]. The existential risks of AI have not been dealt with in this thesis because we believe that if such a prospect were to exist one day, this fear remains today a product of science fiction that bursts into our imaginaries through story telling, and risk diverting resources and attention from the current dangers that this technology raises (Gebru and Torres, 2024).

Nevertheless, the dangers of AI are real, although more permissive. The technology has known numerous hypes and winters since its infancy in the 50s, and we are now at the top of the wave. AI is praised for its achievements and for most people, the future seems bright as AI will be used for the greater good. But behind the hood of these promises, lies a technology that we still cannot fully control and that, if use without care, could accentuate inequalities, social insecurities, systemic racism and sexism, environmental damage, and so on. To account for these harms, the EU needed a regulation.

Even if the AI Act has its flaws, we must recognise that it is a first step in the right direction. The AI Act was first proposed in 2021, long before other legal frameworks were introduced worldwide. It does not come alone and will join the rest of the European legal landscape to strengthen digital rights[63]. At a time when the technology is booming and superpowers such as the US and China are racing to become the leader on the AI market, Europe has decided to take a different approach: to continue to support the EU market, while operating under legal control. Europe has managed to adopt a law on an ever changing technology, which is challenging

[62] The WEF considers that the top four risks in the next ten years are *"extreme weather events"*, *"critical change to Earth systems"*, *"biodiversity loss and ecosystem collapse"* and *"natural resource shortages"*. However, *"adverse outcomes of AI technologies"* still arrives in sixth position, after *"misinformation and disinformation"*.

[63] See Appendix A.1.

to define or understand, in a competitive market where economic incentives drive research and development, and — although only time will tell – without destroying EU businesses and continuing to promote innovation. The task was enormous and regulators did their best to tackle it as best as they could.

As a result, the AI Act appears as a kind of experiment conducted by the European legislator: a "product safety" regulation without a clear defined product and not solely based on safety. Indeed, the AI Act covers a myriad of possible product components and a whole fuzzy family of technologies, and which apply across all industrial sectors. The introduction of fundamental rights into the mix of product safety can also surprise compared to previous regulations, and shows the concerns of the legislator as regard the societal impact of these technologies. These distinctive features will put the NLF and its standardisation system to the test, as such a large scope pushes standards to also be quite broad and incorporate normative elements that they do not usually cover.

The challenge of the AI Act enforcement will also be to contain lobbying efforts from BigTech companies, which now that the AI Act is passed can focus on influencing its implementation. The dominance over the digital market of U.S. based companies such as the GAFAM[64] has raised concerns among the European population (Mönnig and Manouchehri, 2024). The question remains open on whether European law and its enforcement mechanisms can cope with the monetary and economic power of these companies. The story of the AI Act has only just begun, and both experts and the general public will have to wait to see the outcome of this European gamble.

9.7.3 *The legacy of the AI Act and the future of AI regulation*

A complementary approach for AI liability

The AI Act is one of the first legal framework for AI in the world, but it is not the only one, and many will follow. First, even in Europe, other frameworks are meant to complement the AI Act, such as the AI Liability Directive (European Parliament and Council, 2022). Indeed, the AI Act covers *ex ante* accountability of providers, not *ex post* liability. Under the AI Act, a provider can be taken to court for failing to meet compliance obligations when placing an AI system on the European market. But if these obligations are met and a problem still occurs, AI systems currently fall under the recently revised Product Liability Directive (European Parliament and Council, 2024a). To complement these general rules, the AI Liability Directive would apply to fault-based liability claim, which usually requires claimants to prove damage, fault and causality. The proposed AI Liability Directive would create a "presumption of causality" that would give claimants a more reasonable burden of proof (Madiega, 2023). Other complementary measures are also envisaged, such as the disclosure of evidence about high-risk AI systems. Despite the much needed update of this liability regime, some experts still consider it to be insufficient. According to Wachter (2024), limitations include the fact that the damage must be known and tangible, and that claimants must engage in risky and costly procedure.

[64] Google, Amazon, Facebook (Meta), Apple, Microsoft.

Update March 2025
The project of an AI Liability Directive was abandoned by the European institutions after the legislative process appeared to have stalled (Datta and Hartmann, 2025).

International framework: the Council of Europe convention

In parallel to the AI Act, European and extra-European countries were also working on a text with a global reach: the Council of Europe[65] Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law (Council of Europe, 2024a). This Framework Convention, which was opened for signature in September 2024, only a month after the entry into force of the AI Act, is the first-ever international legally binding treaty. Countries can choose to ratify it, but, contrary to the AI Act, it is up to them to determine how to implement its provisions into national laws[66]. The Framework Convention can also apply to the private sector as parties can opt to be directly obliged by the relevant provisions.

As regard to the obligations of the signatories, Chapter III of the Framework Convention sets out general principles to be respected across the life-cycle of an AI system. These principles include for instance "human dignity and individual autonomy", "transparency and oversight", "accountability and responsibility" and "equality and non-discrimination". These principles are similar the "trustworthiness" characteristics of AI systems that correspond to the essential requirements set out in the AI Act (Chapter III, Section 2). However, they are more directly inspired by fundamental rights legislation and rely less on the technical characteristics of AI systems. In addition, the Framework Convention places greater emphasis on the need for digital literacy and skills amongst the whole population, as well as additional safeguards for the rights of disabled people and children. Dignum (2024) summarises that, compared to the AI Act, the Framework Convention has a broader scope, with a strong focus on fundamental rights, and covers more aspects of the AI life-cycle.

Similarly to the AI Act, the Framework Convention is strongly based on risks and will notably require states to carry out risk and impact assessments (Art. 16). However, the "red lines", which in the AI Act are the AI systems that fall into the "unacceptable risk" category and which use is prohibited, are, in the Framework Convention, left to the discretion of the signatory party. Indeed, each party has to *"assess the need for a moratorium or ban"* based on the results of the risk and impact assessments.

The governance mechanisms are also different. Compared to the complex governance structure established by the AI Act[67], the Framework Convention relies on a "Conference of the Parties" to oversee the implementation of the treaty, as well as on international co-operation. The Framework Convention also explicitly requires public consultation and multi-stakeholder involvement in AI governance.

[65] The Council of Europe is an international NGO, not to be confused with the European Council, which is a European institution.

[66] On the contrary, the provisions of the AI Act are directly implementable in all EU Member States.

[67] See Chapter 3, Figure 3.6.

Update March 2025
In December 2024, the Council of Europe released a new tool which provides a clear methodology to carry out risk and impact assessments, as required by the Framework Convention. The tool, called HUDERIA for *"human rights, democracy and the rule of law"*, is a stand-alone, non-mandatory guidance (Council of Europe, 2024b).

Other AI legal frameworks in the world

The AI Act has influenced many countries to follow a similar path and adopt legal frameworks for AI. However, most of these initiatives are still ongoing.

This is the case for instance of Canada's Artificial Intelligence and Data Act (AIDA) (Government of Canada, 2022), which was introduced as part of Bill C-27 in 2022. AIDA takes inspiration from the AI Act as it also introduces a risk-based approach and would be supported by industry standards. On the same side of the Atlantic, the United States have also slowly taken a step towards AI regulation with the Executive Order on the "Safe, Secure and Trustworthy Development and Use of AI" (The White House, 2023). Other legal frameworks are also emerging around the world, such as the British AI Bill (UK Parliament, 2024), the Brazilian AI Regulation (Senado Federal, 2023), the Chinese AI Measures (Cyberspace Administration of China et al., 2023), and various other initiatives[68].

With the AI Act, the European Union hopes to replicate the success of the GDPR and, in particular, the effect it had outside Europe. This phenomenon, known as the "Brussels effect", encompasses both the possibility that non-EU countries and stakeholders will align their production and conduct with European law (the *de facto* Brussels Effect), as well as the possibility that other jurisdictions will adopt similar legal frameworks (the *de jure* Brussels Effect) (Bradford, 2020). If the AI Act does indeed have the potential to have this Brussels Effect, as experts seem to believe (Siegmann and Anderljung, 2022), the legal text and its implementation will not only be key to the success of AI regulation in Europe, but perhaps worldwide.

Update March 2025
AIDA was halted on January 2025 when the resignation of Prime Minister Justin Trudeau and the prorogation of Parliament caused Bill C-27 to die on the order paper (Arai, 2025). However, initiatives could emerge in provincial regulation, such as in Ontario.

In January 2025 in the US, President Donald Trump revoked the Executive Order on AI signed by ex-President Joe Biden (Shepardson, 2025) and signed instead an Executive Order on "Removing Barriers to American Leadership in AI" (The White House, 2025).

9.8 Conclusion

The AI Act was not the beginning of AI regulation at the European level. Indeed, the work of the HLEG started well before and already sets the goals of the European ecosystem: reaching for all high-risk AI systems, a certain level of "trust-

[68] For an overview of state and supra state regulation initiatives worldwide, see (White & Case, n.d.).

worthiness", understood as a the respect for a list of technical properties. The AI Act continues on this notion of trustworthiness by proposing obligations that high-risk systems will have to respect, that are based on technical properties which will be defined in standards, and more precisely in harmonised standards. As of today, the work of CEN-CENELEC JTC 21 focuses mostly on the "AI Trustworthiness framework", a broad horizontal standard meant to tackle all of the essential requirements for high-risk AI systems. However, this broadness and horizontality, inherited from the AI Act and the work of the HLEG, prevents its from defining specific technical requirements which are technology- and context-specific. This is notably true for standards which incorporate ethical and fundamental rights elements, a trend more and more prominent in AI standardisation.

In this discussion chapter, we propose that the three spheres of "trustworthy AI", i.e. ethics, law and technique, should be kept separate. We note that AI standardisation might mark the beginning of a fourth wave of standardisation, where standards deal with increasingly normative questions, though we believe that experts should remain cautious when dealing with these issues. The European standardisation system could be improved and we provide some possible solutions for more transparent processes and a more diverse participation. However, CEN-CENELEC will need to accelerate its process of standard setting if it wants to deliver on time. If hENs are not up to the taste of the Commission, they could be rejected during the assessment process and not have their references published in the OJEU. This strong political choice will have to be made by the Commission, but it would then be acknowledging the partial failure of the current standardisation system. We propose that, in this case, future standards should focus on vertical sectors, where they could go much further into technical detail rather than remaining at an unsatisfactory high level. Given that the successful implementation of the AI Act will probably depend on the effectiveness of its relevant technical standards, the choice of whether or not to reference hENs in the OJEU will be crucial.

Despite its shortcomings, Europe's approach to AI regulation has taken a first step in managing the harms that AI systems can cause to individuals and society as a whole. Only time will tell what its legacy will be and whether these protections will be sufficient to counter the dangers of AI.

Chapter 10
General conclusion

This chapter presents a conclusion to the thesis. We first present our main findings, summarising what we have learned throughout the thesis, examining how we tackled the two research objectives and answering the main research questions.

Research Question – What makes Europe's approach to regulating AI so distinctive from other European regulations?

Research Objective 1 – Understand the interplay between AI ethics charters, regulation and standards.

Research Objective 2 – Comprehend the world of standardisation, the content of standards, their development process and the obstacles they face.

Then, based on these findings, we make our recommendations, primarily aimed at policymakers and standardisation experts, to improve the way AI is regulated in Europe. Finally, we suggest some research directions for future work on the various topics covered by the thesis: the study of ethics charters, the AI Act and technical standards.

10.1 Our findings

10.1.1 Understand the interplay between AI ethics charters, regulation and standards

The story of AI regulation in Europe starts in 2017 with the proposal of a first framework to regulate robots (Chapter 3). If the approach is far removed from what we see today in the AI Act, this proposal highlighted the need for regulation of "intelligent" technologies. At the same time, throughout the 2010s, a wide variety of stakeholders started proposing AI ethics charters and documents (Chapter 2), often listing broad principles, unimplementable as is, but which were understood as the first step towards a more precise and strict means of regulation.

In 2019, the European Commission set up of group of experts to define its own set of guiding principles, known as the High Level Expert Group (HLEG) guidelines, fixated on defining what "trustworthy AI" entailed (HLEG, 2019a). The three components of trust as set out by the HLEG – compliance to ethics, laws and technical requirements – were meant to be distinct and have their separate deliverables in the EU for AI: ethical guidelines, the AI Act and technical standards (Chapter 3).

However, over the years, the boundaries between these three areas have become increasingly blurred. Some guidelines aim to go beyond the ethical sphere, not encouraging critical thinking on the use of technology, but listing vague principles that are widely interpreted by technical teams as means of action (Chapter 2). Some go further, presenting not just values and principles, but characteristics that AI systems – or the processes behind their creation and use – should fulfil in order to be considered "ethical". This is the case, for example, with the HLEG's assessment list (HLEG, 2020a), designed as a checklist enabling companies to test the "trustworthiness" of their AI system, thus displacing this tool into the technical sphere (Chapter 3).

The AI Act itself is strongly inspired by ethics and takes an approach that uses pre-existing ethical guidelines, for example recycling the work of the HLEG in the form of Fundamental Rights Impact Assessments (FRIA) (Chapter 3). However, the AI Act delegates much of the technical work to the standards, relying on them for most of the compliance checks and assessment procedures (Chapter 4). As a result, the AI Act appears to have a double layer: in appearance protecting fundamental rights, but with the means of technical compliance (Chapter 5). In this respect, the AI Act inherits the protective objective of ethical charters, but also constitutes a bridge to the technical world, by opening up to risk management, a practice well known to engineers and technical experts. This way of operating, puts the emphasis on quantitative observations, ignoring the invisible and unmeasurable harms. Therefore, delegating ethical norms to technical standards runs the risk of depriving judges of the normative choice of what is deemed acceptable, entrusting it to technical experts.

On the other hand, while ethics and law always had a complicated relationship, technical standards are supposed to be far removed from their influences: they are "technical" tools, based on science and technique, defining state-of-the-art requirements, guided by the expertise of the people who draft them. Historically, they have been used for product interoperability, encouraging companies to work together to create a single solution that would make people's lives easier. However, in Europe, they play a specific role as a governance and regulatory tool on which the law can rely to specify technical requirements (Chapter 4). Some standards even enjoy a special status under European law, which gives them legal effects such as the presumption of conformity, pursuant to which stakeholders who comply with the standard are automatically deemed compliant with the law.

Additionally to this legal dimension of standards, they might now be dealing with more normative issues. Indeed, the dual objective of the AI Act – to improve the EU internal market by adopting a standards-based compliance framework on the one hand, and to protect the fundamental rights of EU citizens on the other – could mean that standards will now have to address fundamental rights questions (Chapter 5). In addition, ethics is also seen as a topic of interest by standardisation organisations, which are launching numerous initiatives in this area. However, these initiatives have only just begun and, today, there is a clear separation between standards which focus on ethics in general, standards which address several principles of AI ethics but do not go into the details of their practical implementation, and standards which address a single principle, but usually through a narrow technical prism. For instance, the principle of fairness, also often present in AI ethics charters, becomes in the AI Act and, subsequently, in technical standards, a simple measure and mitigation of unfair bias, or even simply a question of checking the "quality' or "representativeness" of data (Chapter 6). The philosophical concept of fairness, strongly linked to moral duty and individual and social behaviour, has therefore become a characteristic of the dataset used to train AI systems that can be verified through various properties.

Nevertheless, the standardisation world fails to recognise that these technical requirements are also strongly value-laden (Chapter 5). Many standardisation experts are not even convinced that ethics and fundamental rights have a place in standards (Chapter 8). Yet, the standardisation community keeps on working on these topics, specifically on "trustworthiness", i.e. ethical criteria translated into a list of verifiable technical properties.

Ethics charters, the AI Act and technical standards do not operate in a vacuum. There is a great deal of interaction between them, as they draw inspiration from each other or rely on each other to function. First, ethics charters take on a more technical role, sometimes becoming checklists of the "ethical" characteristics of a systems. Second, the AI Act is inspired by ethics charters and policy documents, but also plans to rely on the technical requirements defined in standards. Lastly, standards take on a legal role in the EU , drawing on the legal requirements of the AI Act and addressing ethics and fundamental rights questions.

This approach of blurring the boundaries between fields and mixing regulatory mechanisms is encouraged by the EU with its discourse on AI "trustworthiness". However, this discourse does not take into account the specific nature of each of these mechanisms, what they are good at, and their limitations.

10.1.2 Comprehend the world of standardisation, the content of standards, their development process and the obstacles they face

In ethics charters around the world, vague themes are reused inconsistently and interpreted differently by various stakeholders. Fairness, for instance, is no longer a term encountered only in philosophical theories of justice, nor a purely technical notion in information theory and fair distribution. It has become a term used and reused by various communities, inside and outside AI, by ML specialists, legal experts and social scientists, although it covers several realities (Chapter 2). These ethics charters, with their broad and vague principles, have influenced the AI Act in Europe (Chapter 3), such as the obligations for high-risk AI systems which are only "essential requirements", against which providers must demonstrate the compliance of their AI system, but which are broad enough to be interpreted in different ways[1]. In return, the AI Act strongly relies on technical standards to define technical requirements for these legal obligations (Chapter 4). But it is difficult to define precise technical concepts without knowing in which context they will be applied. Standards are therefore forced to remain vague, simply providing guidance applicable to all AI systems (Chapter 6). In Europe, European Standardisation Organisations have decided to follow the lead of the European institutions and propose an "AI Trustworthiness Framework", which will define the basic technical requirements for compliance with the AI Act (Chapter 8). This foundation standard will remain at a very high level, not going too much into technical detail, therefore running the risk of being no more than a minor improvement on ethical charters. The vagueness of ethical principles and the general discourse about AI ethics has therefore spread, from ethical charters to the AI Act and now to technical standards, all the while remaining far removed from practical implementation.

At the same time, European organisations are having difficulty publishing standards. They face a number of obstacles, some of which are inherited from the stan-

[1] For instance: "transparency" (art. 13), "human oversight" (art. 14), "data representativeness" (art. 10) and so on.

dardisation system imposed by the European New Legislative Framework, such as access to standards, sovereignty issues and industry control (Chapter 7). For example, standardisation bodies traditionally sell standards, even harmonised standards – which are supposed to underpin European law. This led to recent case law requiring access without charge to certain harmonised standards, which ended in a half-victory, with the Court recognising the existence of a public interest in free access to standards. However, it is not yet certain that all harmonised standards will be available in the future. At the same time, this case law has initiated a necessary discussion on the copyright of standards that support European legislation, and on the business model of standards bodies. In addition, standards that underpin European law can also come from abroad, with international standards often taking precedence over European ones under the Vienna agreements. Finally, the industrial origin of standards is both their strength, because the experts have knowledge of the field, and their weakness, because they are subject to the strategic goals of companies and to what they wish to include in the standards. The lack of access to standards, the opacity of who draft them, and the interests they defend, call into question the democratic legitimacy of the current system. In the case of AI, these issues are all the more significant as the European Commission requested standards for the AI Act in a hurry. But the current standardisation efforts in Europe still lag behind their international counterparts. This facilitates the lobbying of American BigTech companies, who are pushing for the adoption of standards with less stringent requirements, in particular international standards, which are already ready but quite shallow (Chapter 8).

Alongside the AI trustworthiness framework, standards are being developed for risk and quality management practices for AI. The trend of quality management standards started in the 1980s. Since then, they have enjoyed great commercial success, particularly in the field of social and environmental issues. However, they have also been criticised for failing to deliver on their promise of social progress (Chapter 4). Indeed, these process-based standards do not aim to standardise the properties of the systems themselves, but those of the organisations that develop them. While remaining at the governance level avoids the need for standards to make strong normative choices (Chapter 5), it can also be an open door for industry to do as it pleases. Voices from academia, civil society groups and SMEs have organised to resist pressure to adopt international quality management standards for AI as the basis for the AI Act. They are calling for standards to be more technical and to be developed by AI experts, rather than by people with more expertise in standardisation processes than in AI itself (Chapter 8). But even if there is an incentive to incorporate more technical and stricter requirements into standards, defining these requirements may not be easy, particularly for standards that incorporate issues relating to ethics and fundamental rights (Chapter 5), and are supposed to apply across different use cases and industry sectors (Chapter 9). Until now, standards for AI have remained at the governance level, defining processes that companies must follow, or at the information level, defining terms and methods, without prescribing strict thresholds that systems must meet (Chapters 5 and 6).

Standards remain at a very high level because the AI Act itself is horizontal, as are the ethics charters that preceded it. It is necessary to rethink the structure of the standardisation system, its objectives and the scope of the standards to ensure that they are effectively applicable and that they do not simply recycle the ethical principles and essential requirements of ethical charters and the AI Act.

10.1.3 What makes Europe's approach to regulating AI so distinctive?

Blending technical compliance and normative questions: a challenge for standards

The strategy put in place with the AI Act reflects Europe's desire to blend the approaches of its two main success stories: the regulation of digital rights through a "by design" approach, as in the General Data Protection Regulation (GDPR) (European Parliament and Council, 2016b), and the management of safety risks through product compliance, as in the Medical Devices Regulation (MDR) (European Parliament and Council, 2017b). The mix of risk-based and rights-based regulation was already apparent in the GDPR, but with the AI Act, the European Commission took it to another level by anchoring the regulation in the New Legislative Framework (NLF) for product safety. As a result, the AI Act sets out a dual and sometimes inconsistent discourse, mixing risks and rights, technical compliance and ethical objectives, product safety and fundamental rights, improvement of internal market and protection of individuals.

This duality, however, is new to standards organisations, used to tackle engineering problems. While some fundamental rights questions translate well into technical questions, such as non-discrimination and algorithmic fairness, the link between the two, e.g. what level of bias is deemed acceptable, entails hard normative questions that the technical experts in standardisation groups do not have the competence or legitimacy to define. They therefore found themselves stuck with these impossible questions, refusing (and rightful so) to clearly answer them, choosing instead to produce large open-ended frameworks that the industry will interpret as they see fit.

Normative questions are made ex ante by providers

In product safety frameworks, *ex ante* regulation is the norm. Manufacturers must check the compliance of their products before distributing them on the European market. This *ex ante* approach is usually supplemented by *ex post* regulation whereby manufacturers are held liable in the event of damage. This two-step approach is illustrated by the General Product Safety Regulation (European Parliament and Council, 2023c) for *ex ante* compliance, and by the Product Liability Directive (European Parliament and Council, 2024a) for *ex post* regulation. For AI, this will translate into the AI Act, specifically the obligations of high-risk AI systems, and the AI Liability Directive (European Parliament and Council, 2022). However, there are two key differences between the AI Act and other *ex ante* frameworks: compliance will encompass normative questions, and these questions will be answered directly by AI system providers.

Normative issues, such as the level of acceptability of a risk of discrimination, are generally dealt with judges in courts, as part of an *ex post* assessment, using in particular the principle of proportionality. *Ex ante* assessment, on the other hand, is generally strictly concerned with technical requirements that have an impact on safety, but not on fundamental rights. This is not the case in the AI Act which requires this *ex ante* assessment to consider fundamental rights questions.

In addition, while other NLF regulations rely heavily on notified bodies to carry out conformity assessments, the AI Act leaves much of this assessment to the providers of high-risk AI systems themselves. The answer to "hard normative questions" (Laux et al., 2024), such as which fairness metrics to use in a given context, will therefore be left to AI systems providers. The AI Act therefore differs from other regulations in terms of what is assessed as part of compliance, by whom and at what point in time.

An horizontal approach from beginning to end

AI ethics charters have historically been horizontal, addressing AI issues generally, although more specific frameworks have developed in recent years. Similarly, the AI Act has remained horizontal and aims to cover all AI technologies, even though different systems fall into different risk categories and therefore have different legal obligations. However, within the category of high-risk AI systems, the obligations are the same whether the AI system is a computer vision system as part of a medical device or an Large Language Model used in the education sector. As a result, the standards that are supposed to define the technical requirements for these systems are also designed horizontally, without distinguishing between technologies or sectors. This horizontality is even encouraged by the European institutions, which want standards *"applicable across sectors and systems"* (Soler Garrido et al., 2024).

This horizontality can be understood in ethics charters, the context of an ethical reflection that is supposed to be adapted to each case by moral human agents. The horizontal nature of the AI Act has also been explained by the fact that sector-specific regulations would have been time-consuming to put in place and potentially inconsistent from one sector to another and from one Member State to another. Yet the advantage of a horizontal approach to standards is never clearly explained. Standardisation experts tend to believe that this horizontality of current AI standards is a first step towards more vertical standards. However, in the case of the AI Act, only horizontal standards such as the AI trustworthiness framework will be offered for citation in the Official Journal of the European Union (OJEU) and be granted a presumption of conformity. If vertical standards come later, they will not benefit from the legal and therefore economic advantage of harmonised standards.

The horizontality of standards is inherited from the AI Act, which in turn inherited it from ethics charters. However, while this horizontality worked fairly well in ethics charters, it is challenging in the AI Act and almost impossible to maintain in technical standards.

The influence of the European Commission

Since the 80s, the Commission has entrusted the ESOs to provide standards that will support EU law and has simply been monitoring, at the end of the drafting process, that the standards corresponded to its request. But in recent years, and in particular for AI standards, EU officials have been closely monitoring standardisation efforts, giving guidance to experts when necessary on what is expected of harmonised standards. They have also been more vocal about improving the standardisation system, and particularly the diversity of the participants. The goal of the Commission seems to be to improve the legitimacy of these standardisation organisations, which is all the more necessary as standards deal with normative issues and technologies, such as AI, that receive strong media coverage.

This new interest in AI standards could also stem from a lack of trust in the governance of standards, which are accused of being the mouthpieces of the American BigTech industry. This lack of trust is reflected in particular in the sidelining of ETSI from the standardisation request, as well as the gradual implementation of common specifications in EU regulations, i.e. the Commission's own frameworks to replace harmonised standards in case they fail, and used to put pressure on the ESOs to deliver high-quality standards on time.

If AI standards fail, they could therefore mark the turning point for EU regulation, moving away from the standardisation system inherited from the New Approach.

10.2 Our recommendations

Although the AI Act is a step in the right direction for regulating AI, the European approach can still be improved. In this section, we present our recommendations for European AI regulation, most of which are developed directly in Chapter 9 or, alternatively, result from our discussion throughout the thesis.

10.2.1 On the interplay between ethics, the AI Act and technical standards

Ethics charters should be differentiated from good practices

AI ethics charters present either very broad and abstract elements in the form of inapplicable principles, or, in the more recent documents, very technical elements intended to support possible implementation. If the evolution towards more technical elements is necessary for the correct implementation of the guidelines, it must not be at the cost of restricting ethics to a checklist of technical properties. We therefore suggest that companies commit to more action in favour of AI literacy, AI ethics and critical thinking among their staff, teaching them the importance of their choices in AI development processes rather than giving them a ready-to-use list of elements to check, which risks restricting their personal responsibility.

Future law should avoid mixing risks and rights

The example of the AI Act will set a precedent for European law. If the implementation of the regulation is a success, other regulations relating to digital technologies will probably follow the same model. However, from what we have seen, the AI Act mixes too many frameworks to be truly effective. In particular, the introduction of risks to fundamental rights into a product safety regulation was an attempt to address the specific dangers of AI systems, but is proving difficult to follow for the technical standards supposed to support the legal text. While harmonised standards for the AI Act will have to try and address these issues now that the AI Act is adopted as is, we strongly discourage the use of this type of framework in the future.

Trustworthiness should not be used to discuss solely technical elements

The scope of "trustworthiness" has shrunk from the HLEG guidelines to technical standards, from a three-pillars concept to a simple list of technical properties to be verified. This is due, among other things, to the fact that the definition of trustworthiness in international standards has been transposed into European standards. We believe that when only technical elements are concerned, the term "reliability" should be used instead, to distinguish between the two and maintain a trustworthiness approach that is based on ethics, law and technique.

Normative issues should, wherever possible, be left out of standards

We are seeing a new wave of standards, starting with AI, which is changing the purpose of standards, from technical documents developed by and for industry, to more legislative frameworks drafted by a wide range of stakeholders in order to be more legitimate. However, the European standardisation system fails in its mission to represent the voices of society as a whole because the structure of standardisation organisations are not made to support this goal.

We believe that standards should keep their distance from broad ethics and fundamental rights questions, because despite their best efforts, standardisation bodies do not have the legitimacy to tackle these issues. This is particularly true for horizontal standards, where the definition of ethical criteria or rights in the general case would run the risk of leading to policy capture and ethic washing practices. However, we recognise that for many standards, the definition of technical requirements necessarily implies a normative choice, as in the case of fairness. For these highly normative topics, the definitions of precise criteria should be left to vertical standards. If horizontal standards can create means of *"ethical disclosure by default"* (Laux et al., 2024), they should only present baseline criteria but transparency alone will not provide for compliance with the law.

10.2.2 On the improvement of the standardisation system

Access to standards should be improved

Recent case law has allowed access without charge to certain harmonised standards, but this is, for now, an isolated case. We believe that, as harmonised standards have important legal consequences in the EU, they must be made accessible to all, otherwise they risk undermining the principle of transparency and the right of access to documents. Yet, for this to take effect, discussions are needed on the copyright protection of standards. One potential solution would be for Europe to modify its relationship with the European standardisation organisations (ESOs), by recognising that the standardisation request is a service contract. The European Commission could then invest directly in the standardisation work of the ESOs and allow the resulting standards to be published freely of charge.

The Vienna agreements should be renegotiated

The Vienna and Frankfurt agreements could undermine the EU's sovereignty in standardisation. These agreements therefore need to be renegotiated to make it easier for the ESOs to work on their own standards to support EU law. We acknowledge that this change could have a negative impact on standardisation work, by creating a duplication of initiatives between international and European organisations, which would lead to a proliferation of standards that are sometimes inconsistent with each other. However, we believe that as European law is regional, standards should reflect the particularities of this regional regulatory approach. It is a paradox that Europe creates legislation such as the AI Act in order to produce a global "Brussels Effect" and then open the door to a "California Effect" (Siegmann and Anderljung, 2022) via international standards. In our opinion, standards should emanate from the EU outwards, with a view to influencing international standards, not the other way around, as it is already the case with "traditional" means of regulation, i.e. with European legislation.

Communication efforts around standards should be improved

The lack of manpower in the field of standardisation is at the root of many problems. As standards are currently developed by a handful of stakeholders, their voices necessarily carry more weight, giving greater importance to the voices of those present and delegitimising the process. It is absolutely necessary to step up communication on these issues, whether through awareness-raising campaigns or by incorporating standardisation issues into school curricula.

The development process of standards should be more transparent and more diverse

If standards are to continue to venture into normative questions and have legislative power, the development process of standards needs to be more transparent and more diverse. The ESOs should accommodate to enable greater participation of academic researchers who provide much needed technical expertise. While the European Commission has not yet found a perfect solution for including academics, we believe that a first step could be to recognise individual work on standards. In addition to providing researchers with proof of their involvement, which is necessary for their academic careers, this would also reduce lobbying efforts, since the names of contributors would appear on the standards.

We also suggest reinforcing funding for civil society organisations. To date, only a few organisations receive EU funding to participate. If Europe wants civil society to be as present as the private sector in standardisation, it will have to support it financially. Special measures could also be taken to increase the participation of SMEs and European companies, whether in terms of funding or awareness-raising.

Regulation 1025/2012 should be revised to allow the European institutions to be more resilient in standardisation

A revision of Regulation 1025/2012 on European standardisation could take account of the recent changes made by the AI Act to the New Legislative Framework for product safety in Europe. First of all, this revision should mention the use of harmonised standards for fundamental rights questions, specifying what these standards could address and what their objective is, particularly with regard to the protection of European values. This would help create more certainty about the role of European standardisation organisations in developing normative documents and to avoid political and regulatory capture.

This revision could also be an opportunity to clarify the Commission's role in monitoring standards and the extent to which officials can be involved in steering the work of the ESOs. On this occasion, it could be decided that for standards which are intended to be harmonised and which concern normative issues or European values, the European Commission could have a greater say in the content of these standards, or even be considered as a partner in their development through European specialised agencies. This would reinforce the status of harmonised standards as legal acts as well as their legitimacy, but would also subject them to the same scrutiny as other democratic legislation.

In addition, the revision should incorporate common specifications into the list of technical specifications in Regulation 1025/2012, specifying (1) when they can be used – i.e. under which conditions, (2) what types of frameworks can be adopted – e.g. international standards, frameworks drafted solely by the Commission, etc. – as well as (3) what legislative power they hold – e.g. presumption of conformity, mandatory compliance, etc. As there are not many precedents[2], further clarification is needed for this type of framework to clarify the grey areas.

Finally, a revision of Regulation 1025/2012 could also be the time to really put on the table the option of having a European standardisation agency in charge of either (1) following closely standardisation work in the ESOs, or (2) draft themselves standards and common specifications. Although costly for the EU, this measure would at least enable the Commission to free itself from its heavy dependence on the ESOs and to react quickly in case of problems with standards.

10.2.3 On the AI standardisation strategy

AI standards should address products

According to the Commission, harmonised standards should be *"oriented to AI systems and products"*, as well as *"sufficiently prescriptive and clear"* (Soler Garrido et al., 2024). This goes against the recent trend of developing governance standards, also known as quality management standards, more oriented on processes and very open-ended, in order to be more flexible for the industry. However, we agree here with the Commission that, to support the AI Act, standards should focus in particular on regulating the technology and be technically precise. Otherwise, standards will not be aligned with the legislation which imposes requirements on AI systems directly, rather leaving companies the choice of what general good practice they want to follow. If harmonised standards cited in the OJEU are too generic, providers will have much more leeway to interpret the requirements are they see fit. Harmonised standards will therefore have to define precise terms and criteria that can be technically assessed, rather than mere guidance, if they want to be cited in the OJEU.

AI standards should be refocused on vertical sectors

We propose refocusing standards on vertical sectors rather than remaining at a high, horizontal level where requirements remain shallow and are difficult to define[3]. Although this goes against the will of the Commission who would prefer horizontal standards (Soler Garrido et al., 2024), we believe that refocusing standards on verticals would enable better implementation of the AI Act, with requirements

[2] Only one in the field of medical devices. See Chapter 7, Section 7.9.4.

[3] As discussed in Chapter 9

that are more tailored to the specific context and could therefore go into more technical details. Information standards would then be more accurate about the terms and methods they put forward. And for certain areas where the state of the art is more widely recognised by a large part of the scientific community and has been stable for some time, it would even be possible to consider setting thresholds, as long as they do not concern highly normative choices such as those which affect fundamental rights, e.g. fairness.

Refocusing standards on vertical sectors could also help re-balance the distribution of power within the standardisation bodies, by giving a bigger role to computer science experts and field experts from smaller businesses. Although they are now welcomed in standards organisations, their voices are often overshadowed by large companies, which have more manpower and therefore hold the key to standards. While the development of sector-specific standards will not solve the structural problems of standardisation organisations as regard to industry capture, it will move standards into areas where BigTech companies have less expertise and cannot necessarily make meaningful contributions. SMEs in this sector – focused on particular applications, e.g. AI for medical image recognition – would then have greater legitimacy and would be more listened to, and their needs better taken into account.

The Commission should review AI standards substantially

Scholars tend to think that the Commission does not have the resources to review the content of standards and simply calls on HAS consultants to check whether the scope of harmonised standards matches the scope of the standardisation request and the corresponding regulation (Ebers, 2022). At the same time, the AI Office is recruiting new staff, including technology specialists. With its 140 employees, the AI Office will be a driving force in the implementation of the AI Act. One of its new tasks could be to evaluate standards directly, without going through its usual contractor, Ernst & Young. If a team of computer scientists, engineers, as well as legal and policy experts, could carry out the assessment themselves, the content of the standards could be reviewed and not solely its scope. Because of the high stakes associated with standards, since the AI Act relies on them for many provisions, we strongly encourage this solution, as we believe it will provide better control over the effective implementation of the regulation.

Current horizontal standards should not be cited in the OJEU

While horizontal standards are often seen as a stepping stone for vertical standards, we argue instead that, in the case of AI, they will support the AI Act alone, relegating vertical standards to a secondary role. This risks depriving the AI Act of enforceable means of actions to make providers comply with the obligations it sets forth. Many experts point out that the AI Trustworthiness framework, the CEN-CENELEC JTC 21 broad horizontal standard intended to support several provisions of the AI Act, is not yet complete and could remain at a very high level, which would not allow effective enforcement of the law.

We therefore suggest that this standard, if the claims are verified when the standard is published, should not be cited in the OJEU. Indeed, the legal power granted by the citation is too important to be left to a framework that do not offer proper protection and could create loopholes. The standard can still be adopted by the ESOs and provided to companies as a means of compliance, but it will not grant a presumption of conformity and will make companies more accountable for their conformity assessment procedures.

For future legislation outside AI, the Commission should issue, from the start, one standardisation request for horizontal standards and another for vertical standards. Vertical standards could then be used to provide presumption of conformity for certain categories of products and horizontal standards could provide guidance to cover the remaining gaps.

10.3 Direction for future work

The theme of this thesis – AI regulation – is broad and covers many topics of interest, each of which deserves more attention. However, by covering such a wide range of subjects, we recognise that we have opened many doors. There is still a lot of research to be done on ethical charters, legal acts and technical standards for AI, to close these doors. In this section, we consider various points and directions for future work beyond this thesis.

10.3.1 Working on ethical charters

Although AI ethics charters seem to have been rapidly eclipsed by other means of regulation such as mandatory legislation and technical documents, they are still an interesting object of research. They have inspired these subsequent documents, but they also remain relevant for systems that are not directly covered by legislation or as a complementary approach to mandatory law. With Chapter 2, we have simply scratch the surface of the research possibilities on AI ethics documents. The database we provide[4] can serve as a starting point for future analysis of discourses around AI ethics.

First of all, we know that a term can have different meanings depending on the text in which it is found. For instance, the principle of fairness has permeated different spheres, and is circulated by different actors in a variety of documents[5]. However, it is hard from our initial results to capture the variety of definitions that the principle of fairness encompasses. A research venture could be to use semantic graphs to link the term "fairness" – or any other AI ethics principle – and its related vocabulary, to the context of sentences in which it is used in different documents. This would allow us to validate the results from previous qualitative studies which focus on the various definitions of AI ethics principles used in different documents (Evers, 2024).

Furthermore, the results of our previous study (Gornet and Viard, 2023a) could be validated on this new and bigger database. Such analysis would require to annotate the database with the relative categories of actors from which AI ethics documents emanate – e.g. academia, private sector, public authorities, etc., and to compile our Latent Dirichlet Allocation (LDA) model on the database.

Similarly, a time analysis could be conducted on the database to look at how the principles, vocabularies and general discourses have evolved through time. Coupled with the analysis of the category of stakeholders, the analysis of mesostructures could reveal dynamics, exchanges of ideas, convergence or divergence of discourses. This analysis could also be completed by a mapping of the different actors,

[4] `https://mapaie.telecom-paris.fr/`.

[5] See Chapter 2.

with a graphs of mesostructures – i.e. the institutions which produce these types of documents, and an analysis at micro level, to look at which individuals are present in these institutions, who moves around in different mesostructures and bring with them ideas and vocabularies.

The work initiated on AI ethics documents in this thesis will be continued in a research project called "CHarting Artificial Intelligence (CHAI)", funded by the French National Agency for Research (ANR), and led by Tiphaine Viard[6].

10.3.2 Working on the AI Act

Since the first proposal of the AI Act in 2021 to its final adoption and publication in 2024, the journey of the AI Act has been one to remember. Many have analysed how to interpret provisions, or have emitted critics about the text and its practical implementation. Such studies will likely continue long after the entry into force of the AI Act. The text will be analysed for the years to come and may even become one of the most studied European regulation, alongside the GDPR. As Smuha and Yeung (2024) noted: *"many of the AI Act's core provisions are written in broad, open-ended language, leaving the meaning of key terms uncertain and unresolved"*. As the text continues to apply gradually for the next few years and as sanctions will fall, it will be worth reflecting on how they change the interpretation of the text, similarly to the GDPR and the following case law that completed it and which are essential to the understanding of the regulation, such as the Schrems rulings[7].

What could also be interesting would be to look at the influences of the texts beyond the policy documents and legal texts that we already analysed in Chapter 3, such as how amendments came to be proposed, by whom and for which reasons, as well as how they got included in the text. Indeed, we know that the various stakeholders are lobbying for their interests to be adopted in the regulation. This was particularly apparent in the ethics guidelines of the High Level Expert Group on AI (HLEG), where differences in opinions led to the deletion of key terms such as the term "red lines" (Metzinger, 2019). An example of this for the AI Act would be the term "General Purpose AI" (GPAI) that was first proposed by the EU Council, however the Parliament had another term for it: "foundation models" (European Parliament, 2023a). Yet, the term GPAI seems to appear mostly in BigTech documents (Google, 2021; Microsoft, 2021), while the term foundation models appear in scientific publications from Stanford University (Bommasani et al., 2021). If this is only an example and should not be taken as a rule, the study of how the regulation came to be could nevertheless help unfold the lobbying from both private and public institutions that took place during the negotiations of the AI Act, to make sense of the patchwork of terms and definitions that appear in the final version.

In addition, since the AI Act is a horizontal framework, work will be needed to understand what is required of AI systems in different scenarios. EU affiliates, such as the AI Office or the AI Board, as well as stakeholder groups, will need to provide clarifications on grey areas of the text and guidelines for compliance. However, academic research can also help, by examining the various loopholes that still persist in the text, and propose solutions to address them.

Finally, the EU governance framework for AI could also be of interest to academics who could study how the complex structure established by the AI Act[8] would function in practice. Comitology studies could focus on the interaction between the AI Office and Member States through the AI Board or the market surveil-

[6] https://www.telecom-paris.fr/chart-artificial-intelligence-project.

[7] The Schrems ruling refer to three rulings from the Court of Justice of the European Union (CJEU, 2015, 2020, 2024a) as regard, among other things, to the transfer of personal data from the EU to the US (Schrems I and II) and targeted advertising (Schrems III).

[8] See Chapter 3, Figure 3.6.

lance authorities, as well as between the AI Office and stakeholder groups in the Scientific Panel and Advisory Forum. Academic studies could also help to map the relevant national actors and the different practices for the AI Act implementation in each of the Member States, including market surveillance, regulatory sandboxes and measures for innovation.

10.3.3 Working on standards

It is hard to conduct academic research on standards when the objects of this research are copyrighted and hard to access. Indeed, today, standards are most often behind a paywall, and even when researchers can have access to them, they cannot directly share what is inside. In this thesis, we have avoided being too descriptive about the content of these standards, giving examples where necessary, but generally remaining at a high level of comment. Furthermore, the development of AI standards is only just beginning and we should have more of them to study in the coming years. As a result, it may be easier today, and just as interesting, to focus on the standardisation process rather than on the content of these standards. In this thesis, we have focused in particular on the standardisation of AI and its specific features, but there are still many ways of studying this ecosystem through the sociology and geopolitics of standards.

Legal experts have been slow to take an interest in the development of standards, and further work is still needed to understand the role of standards in EU law and how standards fit into this regulatory framework. The protection of standards by copyright is a topic that we notably expect to take central stage over the next few years, particularly as regards harmonised standards. As regard to the AI Act, it will be interesting to see which standards are published and the impact they will have on the implementation of the legal obligations.

There is also a need to deepen the work on AI standards by linking it to the older academic literature on economics, political science and law. What could be analysed is in which cases actors cooperate on AI standards and in which cases they do not, and why. According to von Ingersleben-Seip (2023), cooperation is what prevails today in ISO/IEC SC 42. But, as we have seen, the history of software standards shows that these spaces can just as easily become battle arenas to defend economic interests[9]. A study on the economic tensions in AI standardisation could therefore uncover hidden dynamics.

In addition, further work could focus on the governance system of standards bodies, in particular their membership status, their voting system and their link with other organisations. For instance, studies could examine how the governance system of standards organisations impacts power dynamics within groups and the subsequent content of standards.

We have also witnessed that there are very few studies on the statistical participation of stakeholders in standardisation organisations. Such studies, similarly to the one we carried out in Chapter 8, are essential to understand which types of stakeholders or countries are interested in standardisation. Such studies could be conducted in other working groups or technical committees, outside of AI, in order to compare participation from one field to another. These studies could be completed by an analysis of the exchange of information within the organisations, to understand who contributes and on which topics. Quantitative studies could also include an analysis of the evolution of hENs in recent years: are there more or

[9] See the example of the OOXML standard in Chapter 7, Section 7.7.2.

fewer hENs published in the OJEU, for which products, are they more guidelines or requirements, are they drafted by CEN-CENELEC or adopted from ISO-IEC standards?

The sociological study of AI standardisation has also just begun. We found only one work in which the author went into the field to study the internal dynamics of an association involved in standardisation activities, namely the Internet Engineering Task Force (IETF), and followed its evolution over time (Cath-Speth, 2021). Other studies should do the same for different standardisation organisations and committees, in order to study the interaction between them, at the meso or macro level. Interesting fieldwork include ISO/IEC SC 42, IEEE or ETSI. The study of CEN-CENELEC JTC 21 should also be continued, taking our study as a basis and developing it further[10]. It could be interesting, for instance, to follow a particular working group or a particular standard and see how the work develops, what specific obstacles are encountered, and so on. Our study could also be completed by a micro-level analysis of the trajectories of the people involved in standardisation, how they became interested in the topic and their interactions with other individuals.

Furthermore, while recent work has examined the reasons for the growing distance between ETSI and the European Commission[11], we believe that further work is needed to understand the interactions between the European institutions and the ESOs. Indeed, as harmonised standards are increasingly used to support EU legislation, the European institutions find themselves more and more reliant on the work of the ESOs. The recent introduction of the possibility for the Commission to draft common specifications in case standards are not ready or do not reach the required level of quality to give presumption of conformity to the legal text, the repeated calls to have more diversity in the standardisation process or to accelerate the delivery of standards, as well as the recent case laws on access to harmonised standards, show the growing tensions between the standardisation organisations and the European institutions. An analysis of the evolving relationship between the European institutions and the ESOs could therefore help to shed light on the current situation.

Finally, a comparison between AI standardisation and other standardisation domains is needed to better understand the specificity of AI standards. This comparison could be done with the sector of medical devices, which is often cited as an example for European regulation. Standards for medical devices have been in place for several years now, they are generally considered to be effective and the NLF seems to work perfectly for this regulation (Mazzini and Scalzo, 2023). A comparison of the two regulations[12], their respective standards and the history of standards setting, could highlight the differences with AI standards in order to find solutions to the obstacles that AI standards are facing today. Another possible comparison could be with ecodesign standards, as the two share many similarities: they are fairly recent and touch upon social concerns. Furthermore, ecodesign regulation has taken a different route than AI, with a horizontal European regulation but an approach to standardisation that mixes horizontal and vertical standardisation (Bundgaard and Huulgaard, 2023). A comparative study could show what lessons can be learned from these experiences and what strategy should be adopted in the future. Another field for comparison could be the standards for cybersecurity, especially harmonised standards for the Cyber Resilience Act (European Parliament

[10] This fieldwork will be continued by Hélène Herman as part of her PhD thesis.

[11] See Chapter 7, Section 7.9.3.

[12] Or three if we include the in vitro medical devices which have their own regulation.

and Council, 2024d). Work on these cybersecurity standards is just beginning, and it will be interesting to follow their development to see whether they face the same issues as AI standards. The comparison could also focus on the role of academic or civil society actors in standardisation, looking at whether this trend is indeed specific to AI or if it can also be witnessed in other sectors.

Generally speaking, further studies should focus on the role of these stakeholders – academics and civil society – in standardisation. Indeed, the question arises as to whether standardisation is the right place for these groups and whether it should indeed try to integrate more diversified profiles, as the Commission seems to desire, or whether standardisation should remain an industrial sphere with specific expertise. Such study should take into account the economic benefits of having purely industrial players involved in standardisation, the efficiency of the development process with these actors, but also the political will to say that standards emanate from society as a whole. The composition and diversity of standard groups will have a direct impact on their legitimacy – input, output and throughput[13] – and the cost and benefits of each approach should be carefully considered.

10.3.4 Other research directions

During my PhD I met many different people and engaged in many different research directions, not all of which are represented in this thesis. This section therefore aims to shed some light on research directions that I could have explored further and that I hope others will pursue after me.

Audit and certification

Standardisation is not the last step in regulating AI. Indeed, it is part of a wider framework, including certification and accreditation, which is sometimes referred to as the *"tripartite standards regime"*[14] (Busch, 2010). Standards are notably used during audits to verify the compliance of an AI system to technical or legal[15] requirements.

The business of certification is quite different to the business of standardisation. It is not about creating common rules and selling a deliverable (the standard), but rather about selling an expert service. In the case of legal compliance to European law, the conformity assessment process, with the delivery of a CE mark, is the only "official" certification. For this conformity assessment procedure, in the AI Act, third-party assessment is only necessary in a few instances. But most of certification services are actually not about compliance to legal requirements, but about compliance to external standards that do not have any legal effects. These external certification schemes, and labels, are often seen by the consumer as "quality marks" to help them differentiate between otherwise similar products.

However, these labels have their limits. While they are not yet highly developed for AI, the case of the eco-labels for sustainable products has shown that there is no clear link between having a label and having a reduced impact on the environment (Horne, 2009). A mapping of all labels for AI could help navigate this space, including "ethical AI" labels, labels for the AI Act[16], "sustainable AI" labels, etc. The economic incentives and marketing practices of different companies selling labels and certification schemes could also be compared.

Since AI standards do not give clear technical requirements, we can anticipate

[13] To understand the differences between these types of legitimacy, see Chapter 7, Section 7.8.3.

[14] See Chapter 4, Section 4.5.4.

[15] In the case of the AI Act, compliance to harmonised standards will lead to compliance to the legal "essential" requirements of Chapter III, Section 2 of the AI Act.

[16] Even though only the CE mark can show compliance to the legislation, some private labels still claim they can help providers to comply with the legislation. This is the case for instance of GoodAlgo (n.d.). Other labels are briefly discussed in Chapter 5, Section 5.3.5.

that normative choices will be pushed to the audit process, down the line. This may result in AI auditors having more flexibility but also more responsibilities in assessing AI systems. However, in order to avoid setting arbitrary criteria, in the absence of strict instructions in standards, they may have to rely on what providers consider to be sufficient. Similarly to the work conducted for Chapter 8, a field study amongst AI auditors could therefore help reveal difficulties and novelties with AI audit compared to the audit of regular products.

During my PhD, I worked closely with the Laboratoire National de Métrologie et d'Essais (LNE)[17]. Although this work did not resulted in any publication, I learned a lot about the evaluation of AI systems and AI certification.

[17] `https://www.lne.fr/en/offers/thematic/articial-intelligence.`

Metrics of explainability

The field of Explainable AI (XAI) has been thriving in recent years due to soaring needs for new explainability methods that help humans better understand the decision of machine learning systems and especially neural networks. Unfortunately, these techniques are often disconnected with the actual needs of stakeholders. It has become essential to list desirable properties that an explanation should meet in order to satisfy those needs. On the other hand, the proliferation of explainability techniques has raised a selection problem: which explanation is the best and on which criteria? Desirable properties based on human judgement are often difficult, if not impossible, to quantify, but others, based on mathematical properties that the system should meet, can be defined. These properties can then be used as quality estimators, or explainability metrics, to evaluate explanations and rank the different explanability methods.

There is a need to bridge the gap between high-level desirable properties and explainability metrics, in order to identify to what extent we can quantify what a "good" explanations is. But it is as equally important to understand which parts of this evaluation should be left to user studies, which are more suited to assess the relevance of using an explanability methods depending on the context (the experience of the recipient of the explanation, the potential dangers of a specific use case, etc.).

During my PhD, I started to analyse academic articles, mostly from the cognitive science field, listing desiderata for explanations. I found that these desiderata could classified into six main axes (*Truthfulness*, *Informativeness*, *Stability*, *Low overload*, *Interpretability* and *Actionability*). A list of these high-level desiderata can be found in Appendix A.6. Further work needs to be carried out to link these desiderata to quantifiable explainability metrics that already exist in the computer science literature. A few of these metrics, with their respective desiderata, are shown in Figure A.4.

Preliminary results show that three desiderata are well covered by existing metrics: *Truthfulness*, *Informativeness* and *Stability*. However, it may not be possible to measure *Interpretability* without user studies. Further work is needed to confirm these preliminary hypotheses. In particular, a systematic review of the literature is required to properly map all existing metrics and identify gaps.

Tailoring AI explanation to the context of use

The need for explainability will depend on the context of use. Main factors could include the level of expertise in AI of the receiver of an explanation, the potential dangers of the technology, its degree of automation, etc.

In a paper, I started to think about what types of explanation could be used for a special use case – AI-based tools to help identify road safety violations such as failure to wear a seatbelt, or use of a phone while driving:

> Gornet, M. and Maxwell, W. (2023a). L'IA explicable appliquée à la détection de ceintures et de téléphones au volant. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA @ PfIA 2023)*, pages 46-56, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). `https://hal.science/hal-04158889/`

In this work, I explain that, for instance, methods such as LIME or SHAP are more suitable for an experienced user, while counterfactual explanations are easier to understand for a layperson. Future work could focus on mapping which type of explainability method could be used in which general context.

Designing fair AI systems

When dealing with machine learning, engineers tend to focus on improving certain aspects of performance of their system, such as efficiency, possibly dismissing other important criteria, like fairness. This mindset can have dreadful consequences for companies as well as for end users and may yield discrimination, for instance when resulting in automated facial recognition systems that work better for white men than for women of color (Buolamwini and Gebru, 2018). Researchers have long reduced fairness to a data issue: if the learning data is unbalanced, the system is quite likely to be biased. But this belief overlooks other parameters or coding choices that are also likely to affect fairness. Which coding choices really affect fairness and what are the trade-offs with efficiency?

Before this thesis, I had started to work on how to design fair AI systems, especially for facial recognition. This lead in a paper where we looked at several design "coding" choices, apart from data, that could impact the fairness of the AI system:

> Gornet, M., Kirchner, C. and Tessier, C. (2022b). Operational fairness when coding facial authentication. `https://hal.science/hal-04447868v1/`

Future work could focus on quantifying the impact on algorithmic fairness of data representativeness compared with other coding choices.

The harms of AI on queer identities

When systems are developed for the masses, divergent identities that do not fit the model are excluded from the technology and therefore from the society that revolves around it. This is even truer for machine learning systems, based on data that, for certain categories of people – including for queer people, does not always exist; which learn patterns that, statistically, only work for larger groups; and whose sole purpose is to fit individuals into boxes, assigning them labels that often do not correspond to queer identities. For the most part, these problems are inherent to the nature of machine learning systems, which clashes with that of queer identities.

I have started to explore these issues in a short paper:

> Gornet, M. and Viard, T. (2023b). Queer identities and Machine Learning. `https://hal.science/hal-04763352`

Future work could consist of a case study on the impact of the use of a certain AI system on queer people, for instance in healthcare.

Role of the AI engineer: competency and responsibility

The dangers posed by AI systems are often due to a combination between intrinsic technical limitations and a misunderstanding of the technology that leads to its misuse. Biases are a case in point: they are not a simple problem to fix because they are inherent to machine learning technologies.

I have already explored in a work the main limitations of machine learning systems:

> Gornet, M. Limitations techniques du machine learning et choix de conception. In Nevejans, N. *Décision humaine, décision de l'IA*, to be published.

Future work could focus on whether the major dangers posed by these technologies imply that AI should be developed or operated by trained individuals with specific, verifiable skills[18].

[18] Sharma (2024) proposes to professionalise AI engineering by requiring AI engineers to obtain licenses to build commercial AI products.

Chapter A
Appendices

A.1 Legal ecosystem of the AI Act (Chapter 3)

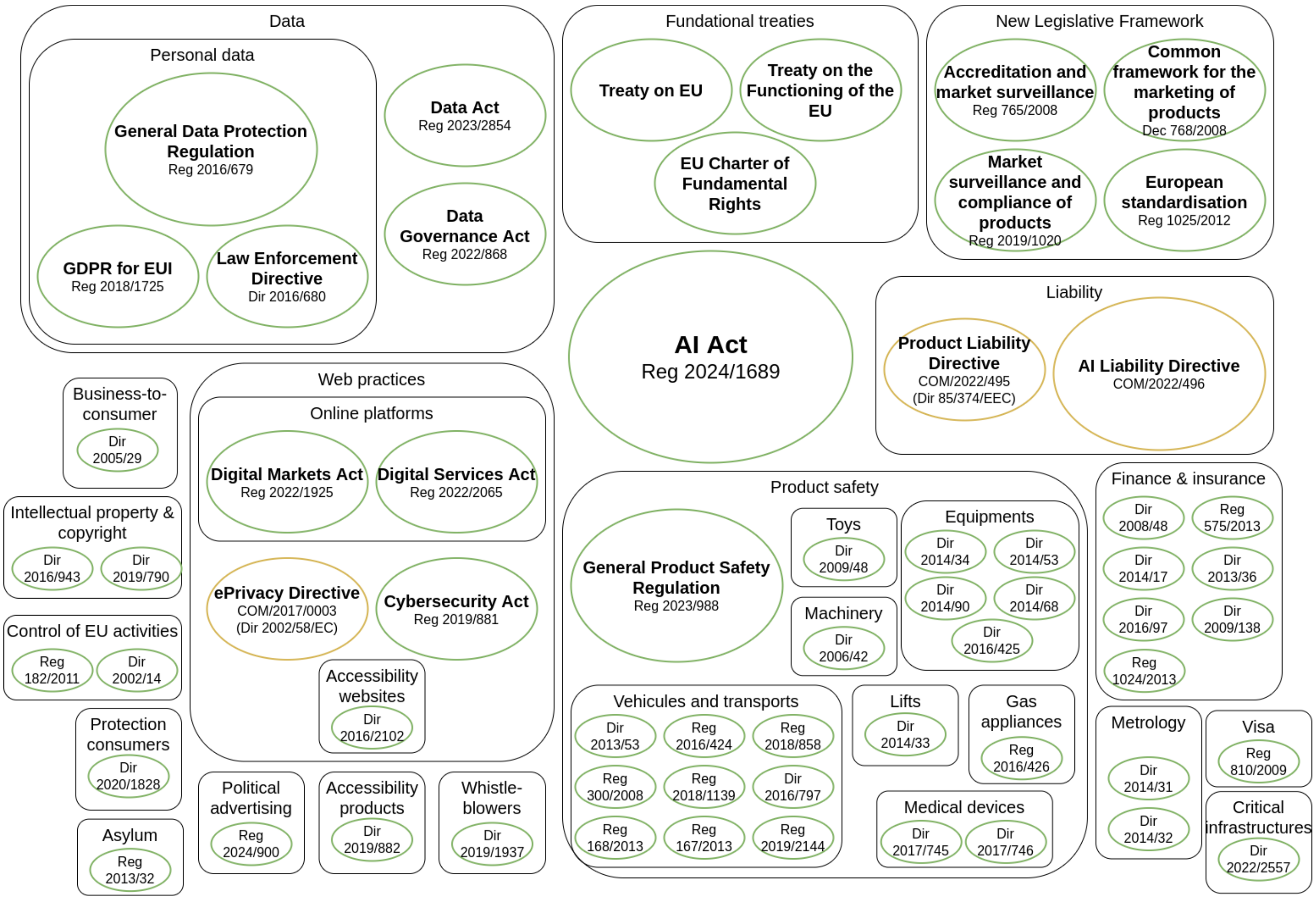


Figure A.1: Legal ecosystem of the AI Act. Each Regulation and Directive that appear is cited in the AI Act, yellow bubbles are texts in construction or revision, green bubbles are already voted. The references to the legal texts are organised by the author.

A.2 List of what the European Commission's delegated acts, guidelines and implementing acts may contain in the AI Act (Chapter 3)

- Delegated acts may be adopted for:
  1. High-risk systems, in order to:
     - amend conditions where an AI system under Annex III is exceptionally not considered high-risk (Art. 6(6) and 6(7))
     - amend Annex III to add, modify or remove AI systems from the list (Art. 7(1) and 7(3))
     - subject high-risk AI systems under Annex III to a conformity assessment with a notified body (Art. 43(6));
     - amend Annex IV on the technical documentation, Annex V on the EU declaration of conformity, Annex VI and VII on conformity assessment (Art. 11(3), 47(5) and 43(5));
  2. for GPAI models, in order to
     - modify the threshold above which GPAI models are considered to present a systematic risk (Art. 51(3));
     - amend Annex XIII to specify or update criteria by which a GPAI model is considered to present a systematic risk (Art. 52(4));
     - amend Annex XI and XII on technical documentation and transparency obligations for GPAI model providers (Art. 53(5) and 53(6)).
- Guidelines may be used to precise:
  - when AI system in Annex III are not high-risk, and provide a comprehensive list of examples for systems that are considered high-risk and those which are not (Art. 6(5));
  - the elements of a quality management system (Art. 63(1));
  - how to apply Article 3 on the definition of high-risk AI systems, Article 5 on prohibited practices, Articles 8 to 15 and 25 on the requirements for high-risk AI systems and the responsibilities across the AI value chain, and Article 50 on transparency obligations (Art. 96(1));
  - the provisions on substantial modification (Art. 96(1));
  - the relation between the AI Act and other harmonised legislation (Art. 96(1)).
- Implementing acts may be used to:
  - approve a code of practice for the transparency obligations of certain AI systems (Art. 50(7)) and for the obligation of GPAI models (Art. 56(6));
  - establish "common specifications" in the absence of adequate harmonised standards (Art. 41(1));
  - establish "common rules" in the absence of adequate codes of conduct (Art. 50(7) and 56(9)).
  - suspend or withdraw the notification of a notified body (Art. 37(4));
  - precise arrangements for regulatory sandboxes (Art. 58(1)) and real-world testing plans (Art. 60(1));
  - establish the scientific panel and precise its procedures (Art. 68(1) and 68(4));
  - precise the fees for access to the pool of experts (Art. 69(2));
  - give a template for post-market monitoring plans (Art. 72(3));
  - detail the conditions of evaluations of GPAI models by the AI Office (Art. 92(6));
  - precise the procedure for fines (Art. 101(6)).

A.3 List of AI fairness standards with their respective annotations (Chapter 6)

Standard name	Horizontal/ Domain	Importance Fairness	Information/ Requirements	Content	Aspect Fairness
ANSI/CTA 2089 Definitions and Characteristics of Artificial Intelligence	Horizontal	2	Information	Vocabulary	Bias
ANSI/CTA 2090 The Use of Artificial Intelligence in Health Care: Trustworthiness	Health Care	2	Requirements	Design	Bias
ANSI/CTA 2096 Guidelines for Developing Trustworthy Artificial Intelligence Systems	Horizontal	2	Information	Vocabulary	Bias
ANSI/CTA 2116 Artificial Intelligence in Health Care: Practices for Identifying and Managing Bias	Health Care	3	Information	Vocabulary	Bias
BS 30440:2023 Validation framework for the use of AI within healthcare	Health Care	1	Requirements	Unknown	Unknown
BS 8611:2016 Guide to the ethical design and application of robots and robotic systems	Horizontal	1	Requirements	Unknown	Unknown
BSI FLEX 236 V1.0:2022-01 Enabling the development of inclusive standards	Horizontal	3	Requirements	Governance	Accessibility, Participation
CAN/CIOSC 101:2019 (R2021): Ethical design and use of automated decision systems	Horizontal	1	Requirements	Governance, Design	Participation, Accessibility, Bias
CEN-CENELEC AI trustworthiness framework	Horizontal	2	Requirements	Governance	Bias
CEN-CENELEC Concepts, measures and requirements for managing bias in AI systems	Horizontal	3	Requirements	Metrics	Bias
CEN-CENELEC Quality and governance of datasets in AI	Horizontal	3	Requirements	Design	Bias
CEN-CENELEC Check List for AI Risks Management	Horizontal	2	Requirements	Governance	Bias
CEN-CENELEC Impact assessment in the context of the EU Fundamental Rights	Horizontal	1	Information	Social context	Unknown
CEN-CENELEC Guidelines on tools for handling ethical issues in AI system life cycle	Horizontal	1	Information	Unknown	Unknown
CEN-CENELEC Guidance for upskilling organisations on AI ethics and social concerns	Horizontal	1	Information	Methods	Unknown
CEN-CENELEC Competence requirements for AI ethicists professionals	Horizontal	1	Requirements	Governance	Unknown
ETSI GR ENI 018 V 2.1.1 Introduction to Artificial Intelligence Mechanisms for Modular Systems	Horizontal	2	Information	Vocabulary	Bias
ETSI GR SAI 004 V 1.1.1 Securing Artificial Intelligence (SAI) – Problem Statement	Horizontal	2	Information	Vocabulary	Bias
IEC SRD 63416:2023 ED1 Ethical considerations of artificial intelligence (AI) in AAL	Health Care	1	Requirements	Governance	Unknown
IEEE P2863 Recommended Practice for Organizational Governance of Artificial Intelligence	Horizontal	2	Requirements	Governance	Bias
IEEE 7003-2024 Standard for algorithmic bias considerations	Horizontal	3	Requirements	Governance, Design	Bias
IEEE 7000-2021 Standard Model Process for Addressing Ethical Concerns during System Design	Horizontal	1	Requirements	Governance	Participation, Accessibility, Bias
IEEE 7007-2021 Ontological Standard for Ethically Driven Robotics and Automation Systems	Horizontal	1	Information	Ontology	Bias
ISO/IEC TS 12791:2024 Treatment of unwanted bias in classification and regression machine learning tasks	Horizontal	3	Information	Methods	Bias
ISO/IEC 19795-10:2024 Quantifying biometric system performance variation across demographic groups	Biometrics	3	Information	Metrics	Bias
ISO/IEC TR 22116:2021 A study of the differential impact of demographic factors in biometric recognition	Biometrics	3	Information	Social context	Bias
ISO/IEC TR 21221 Beneficial AI systems	Horizontal	1	Information	Social context	Unknown
ISO/IEC TS 22443 Guidance on addressing societal concerns and ethical considerations	Horizontal	1	Requirements	Governance	Unknown
ISO/IEC 22989:2022 Artificial intelligence concepts and terminology	Horizontal	2	Information	Social context	Bias, Participation
ISO/IEC 23894:2023 Guidance on risk management	Horizontal	2	Requirements	Governance	Unknown
ISO/IEC TR 24027:2021 Bias in AI systems and AI aided decision making	Horizontal	3	Information	Metrics	Bias
ISO/IEC TR 24028:2020 Overview of trustworthiness in artificial intelligence	Horizontal	2	Information	Vocabulary	Bias
ISO/IEC TR 24368:2022 Overview of ethical and societal concerns	Horizontal	1	Information	Social context	Unknown
ISO/IEC 24714:2023 Cross-jurisdictional and societal aspects of biometrics	Biometrics	2	Information	Social context	Accessibility
ISO/IEC 42001:2023 Artificial intelligence Management system	Horizontal	2	Requirements	Governance	Bias, Accessibility
ISO/IEC 42005 AI system impact assessment	Horizontal	2	Requirements	Governance	Unknown
LNE Certification standard of processes for AI	Horizontal	1	Requirements	Design, Governance	Participation, Bias
Microsoft Responsible AI Standard v2	Horizontal	3	Requirements	Design, Governance	Bias, Participation, Accessibility
NIST Artifical Intelligence Risk Management Framework	Horizontal	3	Information	Vocabulary	Bias
NIST Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials	Biometrics	3	Information	Metrics	Bias
NIST SP 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence	Horizontal	3	Information	Vocabulary	Bias
VDE SPEC 90012 V1.0 VCIO based description of systems for AI trustworthiness characterisation	Horizontal	2	Requirements	Design, Governance	Bias, Participation

Table A.1: List of AI fairness standards with their respective annotations.

A.4 Grid of themes for interviews (Chapter 8)

Each sub-theme – in **bold**, was addressed in all the interviews, but the questions were more flexible. The questions are given here as examples only. The sub-themes are grouped in more general themes – in *italic*, that were used to build the sections of Chapter 8 and organise the discussion.

The expert's path in standardisation

Background. Ex: Could you introduce yourself? What is your background? Who do you work for?

Standardisation groups. Ex: Can you give us the names of the standards you are working on or have worked on at CEN-CENELEC, ISO/IEC and other organisations? What is/was your role (conevnor, editor, etc.)?

Reasons for joining. Ex: Can you tell us how you got involved in standardisation? Why did you decide to get involved?

Organisation of work

Meetings. Ex: Can you tell us about how the standardisation work is organised? How often and for how long do you meet? How do you decide which subjects to cover and which elements to include in the standard?

NSBs (for experts outside of France). Ex: Could you tell us about the organisation within your NSB? How do you get in? What types of experts are involved? How often do you meet?

European Commission presence. Ex: Have you interacted with EU officials? Do you see them in meetings? Does their presence help or hinder standardisation efforts?

The content of AI standards

Type of content. Ex: For the standards you are working on, can you tell us what type of elements they contain? Are they definitions, metrics, requirements, etc.? How precise are they?

Ethics. Ex: Can ethics and fundamental rights be standardised? If so, how?

Specificity of AI. Ex: Is there anything special about AI standardisation? If you have worked on other topics, how do they compare?

Strategy

Alliances. Ex: What other experts do you usually work with or talk to? Do you form groups or alliances?

Company policy. Ex: How many experts from your organisation/company/institute are involved in AI standardisation? How do you organised?

Difficulties and solutions

Personal experience Ex: What was positive about your experience in standardisation? What was negative? Were there any frustrations?

Changes and improvements Ex: If you could change anything in the standardisation system, what would you do? Do you it is feasible in practice?

A.5 Standardisation experts' backgrounds and affiliations (Chapter 8)

ID	Affiliation	Background
P1	Institute	Governance
P2	Consulting	Humanities
P3	BigTech	Computer science
P4	Consulting	Governance
P5	Academia	Computer science
P6	Consumers	Humanities
P7	Academia	Humanities
P8	Institute	Computer science
P9	SME	Computer science
P10	Consulting	Humanities
P11	BigTech	Governance
P12	BigTech	Governance
P13	SME	Computer science
P14	Corporation	Governance
P15	Institute	Computer science
P16	Academia	Humanities

Table A.2: Table of interviewed experts' backgrounds and main affiliations.

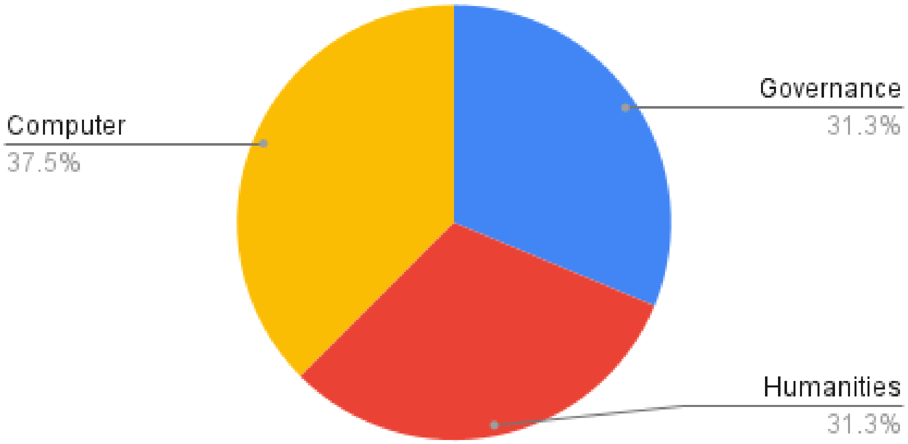


Figure A.2: Pie chart of interviewed experts' background.

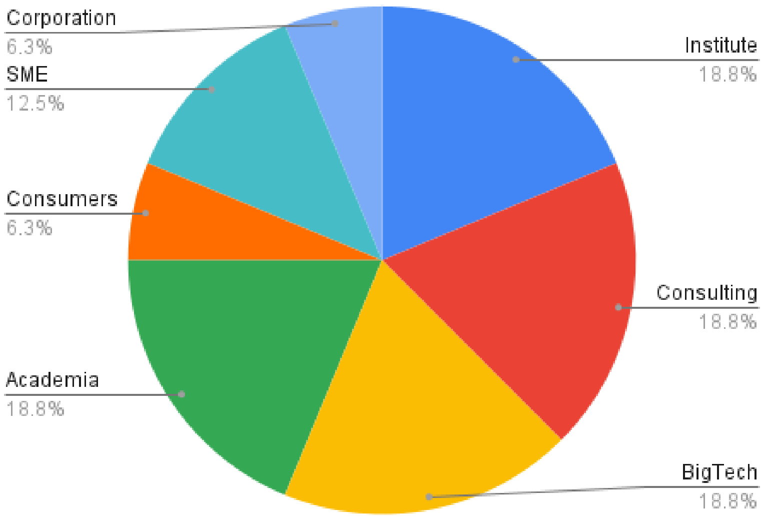


Figure A.3: Pie chart of interviewed experts' main affiliation.

A.6 Metrics of explainability (Chapter 10)

List of desiderata for explanations

- ***Truthfulness*** – The explanation corresponds to reality of the model's behavior, input and ground truth:
  - *Faithfulness* (synonyms: Importance, Representativity, Priorization, Proximity, Certainty, Coherency, Consistency, Similarity, Adequacy, Contextfulness, Necessity, Relevant): the explanation is in line with the model's prediction. Means: by perturbation of important features;
  - *Fidelity* (synonyms: Invariance, Differenciable): the explanation is in line with the input. Means: perturbation of the input;
  - *Soundness* (synonyms: Correctness, Realistic, Exactness, Consistent, Quality): the explanation is in line with ground truth. Means: by localisation (of a mask for instance).
- ***Informativeness*** – The explanation gives all the relevant reasons, and only those reasons, for a prediction:
  - *Conciseness* (synonyms: Compactness, Parcimony, Reasonable, Selective, Simplicity, Necessity): the explanation is short and on point;
  - *Completeness* (synonyms: Broadness, Breadth, Aspect Coverage, Complexity, Sparcity, Sufficiency): the explanation covers all necessary aspects of the prediction;
  - *Generalizability* (synonyms: Generality, Scalability, Transferability): the explanation can be used for other predictions.
- ***Stability*** – The explanation does not change when an insignificant change is made:
  - *Robustness* (synonyms: Reliability, Minimum perturbation): to noise;
  - *(In)sensitivity* (synonyms: Invariance): to input change;
  - *(In)sensitivity*: to output change;
  - *Randomisation*.
- ***Low overload*** – the explanation is not too costly to compute:
  - *Time*;
  - *Performance*;
  - *Others*.

- ***Interpretability*** – the explanation fits human understanding:
  - *Clarity* (synonyms: Transparency, Unambiguity);
  - *Understandable*;
  - *Meaningful*.
- ***Actionability*** – the explanation can be used in practice.

Visualisation of desiderata for explanations and some related metrics

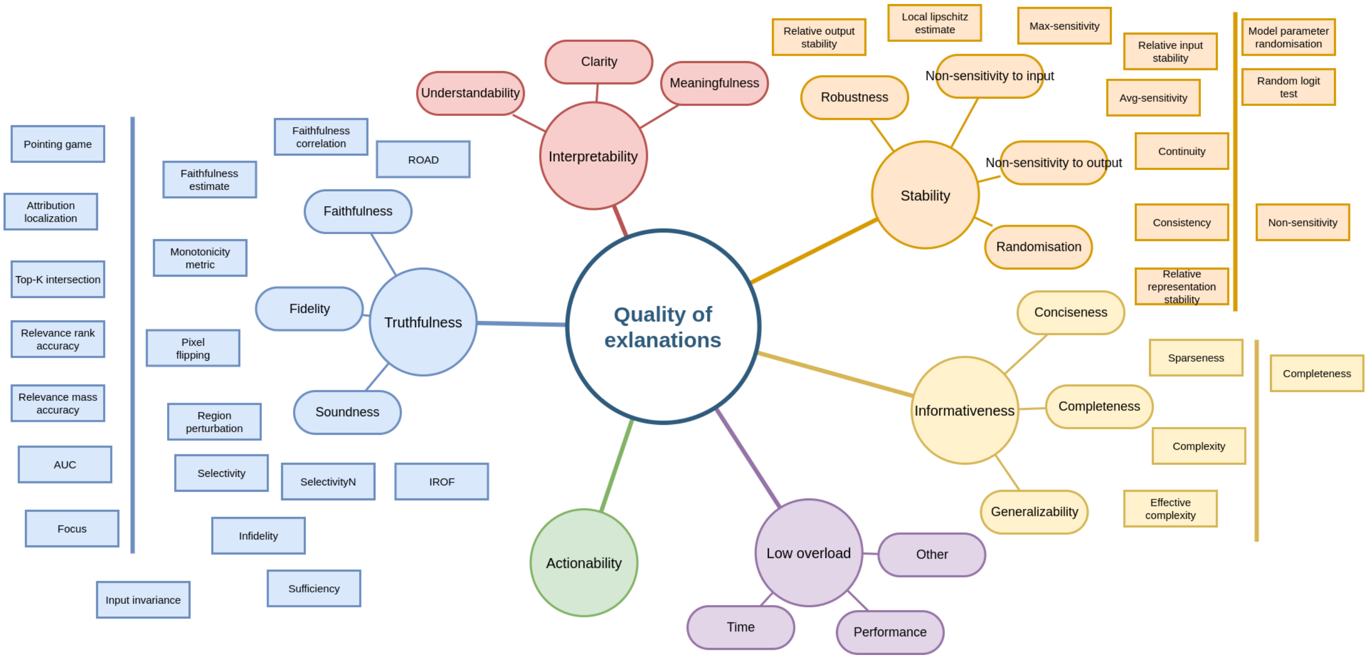


Figure A.4: High-level desiderata for explainability of AI and quantifiable metrics. High-level desiderata identified from the cognitive science literature are in round and ellipse shape, while quantifiable metrics identified in the computer science literature are in rectangular shapes.

References

Abbott, K. W. and Snidal, D. (2000). Hard and Soft Law in International Governance. *International Organization*, 54(3):421–456. Cambridge University Press. https://doi.org/10.1162/002081800551280

Abbott, K. W. and Snidal, D. (2001). International 'Standards' and International Governance. *Journal of European Public Policy*, 8(3). https://papers.ssrn.com/abstract=1550889

Abbott, K. W. and Snidal, D. (2009). The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State. In Mattli, W. and Woods, N., editors, *The Politics of Global Regulation*, pages 44–88. Princeton University Press. https://doi.org/10.1515/9781400830732.44

Abdalkareem, Z. A., Amir, A., Al-Betar, M. A., Ekhan, P., and Hammouri, A. I. (2021). Healthcare scheduling in optimization context: a review. *Health and Technology*, 11(3):445–469. https://doi.org/10.1007/s12553-021-00547-5

Accenture (2016). Universal principles of data ethics: 12 guidelines for developing ethics codes. http://dl.icdst.org/pdfs/files4/62b139fc418fa242173769e431c24623.pdf

Ada Lovelace Institute (2021). Three proposals to strengthen the EU Artificial Intelligence Act. https://www.adalovelaceinstitute.org/blog/three-proposals-strengthen-eu-artificial-intelligence-act/

AFNOR (2016). Vade-mecum: Acteurs du système français de normalisation. https://www.francenormalisation.fr/wp-content/uploads/2016/06/vademecum-normalisation.pdf

AI Ethics Lab (2020). Toolbox: Dynamics of AI Principles. https://aiethicslab.com/big-picture/

AI Standards Hub (n.d.). Standards Database. https://aistandardshub.org/ai-standards-search/

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019). Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 161–170. PMLR. https://proceedings.mlr.press/v97/aivodji19a.html

Albiero, V., Zhang, K., and Bowyer, K. W. (2020). How Does Gender Balance In Training Data Affect Face Recognition Accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. https://ieeexplore.ieee.org/document/9304924

Alexander, M. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press. https://newjimcrow.com/

Alexy, R. (2010). A Theory of Constitutional Rights. *Oxford University Press*. https://global.oup.com/academic/product/a-theory-of-constitutional-rights-9780199584239

Algorithm Watch (n.d.). AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/

Allen, R. H. and Sriram, R. D. (2000). The Role of Standards in Innovation. *Technological Forecasting and Social Change*, 64(2):171–181. https://doi.org/10.1016/S0040-1625(99)00104-3

Almada, M. and Petit, N. (2023). The EU AI act : a medley of product safety and fundamental rights? Technical report, European University Institute. https://hdl.handle.net/1814/75982

Anderson, E. (2020). Controversial Detroit facial recognition got him arrested for a crime he didn't commit. *Detroit Free Press*. https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/

ANEC (n.d.). FAQ & Useful Links. ANEC: The European consumer voice in standardisation. https://www.anec.eu/about-anec/faq-useful-links

ANFR (n.d.). Les résultats des mesures de DAS. Agence Nationale des Fréquences. https://www.anfr.fr/maitriser/equipements-radioelectriques/le-debit-dabsorption-specifique-das/les-resultats-des-mesures-de-das

Angwin, J., Larson, J., Kirchner, L., and Mattu, S. (2017). Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk. *ProPublica*. https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

appliedAI (2023). AI Act: Risk Classification of AI Systems from a Practical Perspective. Technical report, Initiative for applied artificial intelligence. https://www.appliedai.de/en/insights/ai-act-risk-classification-of-ai-systems-from-a-practical-perspective

Arai, M. (2025). What's Next After AIDA? *Schwartz Reisman Institute for Technology and Society*. https://srinstitute.utoronto.ca/news/whats-next-for-aida

Arborus (n.d.). Le GEEIS, Processus de labellisation pour l'égalité, la diversité, l'inclusion. https://arborus.org/label/

Attard-Frost, B., De los Ríos, A., and Walters, D. R. (2022). The ethics of ai business practices: a review of 47 ai ethics guidelines. *AI and Ethics*, 3:389–406. Springer. https://doi.org/10.1007/s43681-022-00156-6

Australia, S. (2006). Standards and the economy. Technical report, Centre for International Economics. https://www.pc.gov.au/inquiries/completed/standards/submissions/standards_australia_cie/subdr133.pdf

avanti europe (2024). All You Need to Know about Common Specifications. https://www.avanti-europe.ch/what-you-need-to-know-about-common-specifications/

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, 253(1):1–13. https://doi.org/10.1016/j.ejor.2015.12.023

Axionable (2021). IA performante et éthique : Axionable décroche la 1ère certification délivrée par le LNE. https://www.axionable.com/ia-performante-et-ethique-axionable-decroche-la-1ere-certification-delivree-par-le-lne-2/

Baeva, D. G., Puntschuh, M., and Binder, M. (2023). Power to the standards: Expert consultation on the role of norms and standards in the European regulation of artificial intelligence (White paper). The Center for Trustworthy Artificial Intelligence. https://www.zvki.de/zvki-exklusiv/fachinformationen/zvki-whitepaper-power-to-the-standards

Balahur, A., Jenet, A., Hupont Torres, I., Charisi, V., Ganesh, A., Griesinger, C. B., Maurer, P., Mian, L., Salvi, M., Scalzo, S., Soler Garrido, J., Taucer, F., and Tolan, S. (2022). Data quality requirements for inclusive, non-biased and trustworthy AI: putting science into standards. Technical report, European Commission Joint Research Centre. Publications Office of the European Union. https://dx.doi.org/10.2760/365479

Balayn, A., Yurrita, M., Yang, J., and Gadiraju, U. (2023). " fairness toolkits, a checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, page 482–495. Association for Computing Machinery. https://doi.org/10.1145/3600211.3604674

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics. https://aclanthology.org/W13-2322.pdf

Baram, M. and Bieder, C. (2022). Standardization and Risk Regulation for High-Hazard Industries. In Laroche, H., Bieder, C., and Villena-López, J., editors, *Managing Future Challenges for Safety: Demographic Change, Digitalisation and Complexity in the 2030s*, pages 85–93. Springer International Publishing. SpringerBriefs in Applied Sciences and Technology. https://doi.org/10.1007/978-3-031-07805-7_11

Barocas, S., Hardt, M., and Narayanan, A. (2021). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. http://www.fairmlbook.org

Baron, J., Ménière, Y., and Pohlmann, T. (2014). Standards, consortia, and innovation. *International Journal of Industrial Organization*, 36:22–35. https://doi.org/10.1016/j.ijindorg.2014.05.004

BBC News (2015). Google apologises for Photos app's racist blunder. https://www.bbc.com/news/technology-33347866

Beck, U. (1992). Modern society as a risk society. In *The Culture and Power of Knowledge. Inquiries into Contemporary Societies*, pages 199–214. https://doi.org/10.1515/9783110847765.199

Becker, H. S. (1976). Art worlds and social types. *American behavioral scientist*, 19(6):703–718. SAGE Publications. https://doi.org/10.1177/000276427601900603

Bellman, R. E. (1961). *Adaptive Control Processes – A Guided Tour*. Princeton University Press. https://doi.org/10.1515/9781400874668

Benbouzid, B. (2023). Fairness in machine learning from the perspective of sociology of statistics: How machine learning is becoming scientific by turning its back on metrological realism. In *FAccT'23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 35–43. Association for Computing Machinery. https://doi.org/10.1145/3593013.3593974

Benbouzid, B., Meneceur, Y., and Smuha, N. A. (2022). Four shades of ai regulation. *Réseaux*, 232233(2):29–64. La Découverte. Translated by Carey-Libbrecht, L. https://shs.cairn.info/journal-reseaux-2022-2-page-29

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., and Williams-King, D. (2025). Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? *arXiv preprint arXiv:2502.15657*. https://arxiv.org/abs/2502.15657

Bernstein, S. and Hannah, E. (2008). Non-State Global Standard Setting and the WTO: Legitimacy and the Need for Regulatory Space. *Journal of International Economic Law*, 11(3):575–608. https://doi.org/10.1093/jiel/jgn022

Bertrand, A. (2024). *Misplaced trust in AI: the explanation paradox and the human-centric path. A characterisation of the cognitive challenges to appropriately trust algorithmic decisions and applications in the financial sector*. PhD thesis. https://theses.hal.science/tel-04661844

Bertuzzi, L. (2022a). AI Act: EU Parliament's discussions heat up over facial recognition, scope. *Euractiv*. https://www.euractiv.com/section/digital/news/ai-act-eu-parliaments-discussions-heat-up-over-facial-recognition-scope/

Bertuzzi, L. (2022b). Commission leaves European standardisation body out of AI standard-setting. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/commission-leaves-european-standardisation-body-out-of-ai-standard-setting/

Bertuzzi, L. (2023). EU's AI Act negotiations hit the brakes over foundation models. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/

Bertuzzi, L. (2024). EU countries make significant changes to AI Board's procedures, structure. *MLex Market Insight*. https://mlexmarketinsight.com/news/insight/eu-countries-make-significant-changes-to-ai-board-s-procedures-structure

Best, E. (2003). Alternative Regulations or Complementary Methods? Evolving Options in European Governance. *EIPASCOPE*, (1):1–10. http://aei.pitt.edu/817/

Bevilacqua, D. (2020). Global Food Safety Regulation and the Interplay between Global Standards and WTO Law: How to Close the Legitimacy Gap? In Eliantonio, M. and Cauffman, C., editors, *The Legitimacy of Standardisation as a Regulatory Technique*. Edward Elgar Publishing. https://doi.org/10.4337/9781789902952.00018

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 210–219. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372860

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 149–159. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/binns18a.html

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2022). The Values Encoded in Machine Learning Research. In *FAccT'22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184. Association for Computing Machinery. https://doi.org/10.1145/3531146.3533083

Bitkom (2016). Germany – Excellence in Big Data. https://www.bitkom.org/sites/main/files/file/import/Germany-Excellence-in-Big-Data.pdf

Bjørnskov, C. (2017). Social Trust and Economic Growth. In Uslaner, E. M., editor, *The Oxford Handbook of Social and Political Trust*, pages 535–556. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.24

Black, J. (2002). Critical reflections on regulation. CARR Discussion Papers (DP 4). ESRC Centre for Analysis of Risk and Regulation, London, UK. http://eprints.lse.ac.uk/id/eprint/35985

Black, J. (2010). Risk-based Regulation: Choices, Practices and Lessons Being Learnt. In *Risk and Regulatory Policy: Improving the Governance of Risk*, pages 185–224. OECD Reviews of Regulatory Reform. https://doi.org/10.1787/9789264082939-11-en

Blind, K. (2004). *The Economics of Standards*. Edward Elgar Publishing. https://www.e-elgar.com/shop/gbp/the-economics-of-standards-9781843767930.html

Blind, K. (2011). An economic analysis of standards competition: The example of the ISO ODF and OOXML standards. *Telecommunications Policy*, 35(4):373–381. https://doi.org/10.1016/j.telpol.2011.02.007

Blind, K. (2016). The impact of standardisation and standards on innovation. In Edler, J., Cunningham, P., Gök, A., and Shapira, P., editors, *Handbook of Innovation Policy Impact*, pages 423–449. Edward Elgar Publishing. https://doi.org/10.4337/9781784711856.00021

Blind, K., Jungmittag, D. A., and Mangelsdorf, D. A. (2012). The Economic Benefits of Standardization – An update of the study carried out by DIN in 2000. Technical report, DIN. https://www.din.de/resource/blob/89552/68849fab0eeeaafb56c5a3ffee9959c5/economic-benefits-of-standardization-en-data.pdf

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008. IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., and Winfield, A. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, 29(2):124–129. https://doi.org/10.1080/09540091.2016.1271400

Bogucki, A., Engler, A., Perarnaud, C., and Renda, A. (2022). The AI Act and emerging EU digital acquis: Overlaps, gaps and inconsistencies. Technical report, CEPS. https://www.ceps.eu/ceps-publications/the-ai-act-and-emerging-eu-digital-acquis/

Bolte, L., Vandemeulebroucke, T., and van Wynsberghe, A. (2022). From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI. *Sustainability*, 14(8):4472. https://doi.org/10.3390/su14084472

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., and Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599. https://doi.org/10.1016/j.neucom.2024.128096

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258v3*. https://arxiv.org/abs/2108.07258v3

Borovicka, T., Jr, M. J., Kordik, P., Jirina, M., Borovicka, T., Jr, M. J., Kordik, P., and Jirina, M. (2012). Selecting Representative Data Sets. In *Advances in Data Mining Knowledge Discovery and Applications*. IntechOpen. https://www.intechopen.com/chapters/39037

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. https://doi.org/10.1093/oso/9780190088583.001.0001

Brand, J. L. M. (2022). Why reciprocity prohibits autonomous weapons systems in war. *AI and Ethics*, 3(2):619–624. https://doi.org/10.1007/s43681-022-00193-1

Bringas Colmenarejo, A., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., and Kinder-Kurlanda, K. (2022). Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI

Regulation. In *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 107–118. Association for Computing Machinery. https://doi.org/10.1145/3514094.3534158

Brunsson, N. and Jacobsson, B. (2002). *A World of Standards*. Oxford University Press. https://global.oup.com/academic/product/a-world-of-standards-9780199256952

Bryson, J. and Winfield, A. (2017). Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer*, 50(5):116–119. IEEE. https://ieeexplore.ieee.org/document/7924235

Bryson, J. J. (2022). Belgian and Flemish Policy Makers' Guide to AI Regulation. Kenniscentrum Data Maatschappij and KU Leuven CiTiP. https://data-en-maatschappij.ai/en/publications/paper-belgian-and-flemish-policy-makers-guide-to-ai-regulation

Bundgaard, A. M. and Huulgaard, R. D. (2023). The role of standards in support of material efficiency requirements under the Ecodesign Directive. *Journal of Cleaner Production*, 385(135599). https://doi.org/10.1016/j.jclepro.2022.135599

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Burden, H. and Stenberg, S. (2022). Regulating Trust – An Ongoing Analysis of the AI Act. Technical report, RISE Research Institutes of Sweden. https://www.ri.se/en/regulating-trust-an-ongoing-analysis-of-the-ai-act

Bureau of Cyberspace and Digital Policy (2024). Risk Management Profile for AI and Human Rights. U.S. Department of State. https://www.state.gov/risk-management-profile-for-ai-and-human-rights/

Burri, M. and Zihlmann, Z. (2023). The EU Cyber Resilience Act – An Appraisal and Contextualization. *Zeitschrift für Europarecht (EuZ)*. https://ssrn.com/abstract=4375552

Busch, L. (2010). Standards, Law, and Governance. *Journal of Rural Social Sciences*, 25(3). https://egrove.olemiss.edu/jrss/vol25/iss3/4

Busch, L. (2011). *Standards: Recipes for Reality*. The MIT Press. https://doi.org/10.7551/mitpress/8962.001.0001

Buyl, M. and De Bie, T. (2024). Inherent Limitations of AI Fairness. *Communications of the ACM*, 67(2):48–55. https://doi.org/10.1145/3624700

Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., and Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI & SOCIETY*, 38. Springer. https://doi.org/10.1007/s00146-021-01380-0

Büthe, T. and Mattli, W. (2010). International Standards and Standard-Setting Bodies. In Coen, D., Grant, W., and Wilson, G., editors, *The Oxford Handbook of Business and Government*, pages 440–471. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199214273.003.0020

Büthe, T. and Mattli, W. (2011). *The new global rulers: the privatization of regulation in the world economy*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691144795/the-new-global-rulers

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, 32:196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Calvi, A. and Kotzinos, D. (2023). Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. In *FAccT'23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1229–1245. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594076

Cambridge Dictionary (2024). Fairness. In *Cambridge English Dictionary Online*. Cambridge University Press. https://dictionary.cambridge.org/dictionary/english/fairness

Cambridge Dictionary (n.d.). Legitimacy. https://dictionary.cambridge.org/dictionary/english/legitimacy

Cameron, J. (1984). The Terminator. [Film].

Cantero Gamito, M. (2018). Europeanization through Standardization: ICT and Telecommunications. *Yearbook of European Law*, 37:395–423. https://doi.org/10.1093/yel/yey018

Cantero Gamito, M. and Marsden, C. T. (2024). Artificial intelligence co-regulation? The role of standards in the EU AI Act. *International Journal of Law and Information Technology*, 32(1). https://doi.org/10.1093/ijlit/eaae011

Cardon, D., Cointet, J.-P., and Mazières, A. (2018). Neurons spike back – The invention of inductive machines and the artificial intelligence controversy. *Réseaux*, 211(5):173–220. La Découverte. Translated by Carey-Libbrecht, L. https://shs.cairn.info/journal-reseaux-2018-5-page-173

Castets-Renard, C. and Besse, P. (2022). Ex ante Accountability of the AI Act: Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance. In Castets-Renard, C. and Eynard, J., editors, *Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime. Comparative Law Perspectives*. Bruylant. https://papers.ssrn.com/abstract=4203925

Cath-Speth, C. J. N. (2021). *Changing minds and machines: a case study of human rights advocacy in the Internet Engineering Task Force (IETF)*. PhD thesis, University of Oxford. https://ora.ox.ac.uk/objects/uuid:9b844ffb-d5bb-4388-bb2f-305ddedb8939

Cauffman, C. and Gérardy, M. (2020). Competition law as a tool to ensure the legitimacy of standard-setting by European standardisation organisations? In Eliantonio, M. and Cauffman, C., editors, *The legitimacy of standardisation as a regulatory technique*, pages 64–90. Edward Elgar Publishing. https://doi.org/10.4337/9781789902952.00009

Celeste, E. (2019). Digital constitutionalism: a new systematic theorisation. *International Review of Law, Computers & Technology*, 33(1):76–99. https://doi.org/10.1080/13600869.2019.1562604

Cellucci, T. A. (2008). Developing Operational Requirements - A Guide to the Cost-Effective and Efficient Communication of Needs. Technical report, U.S. Department of Homeland Security. https://www.dhs.gov/xlibrary/assets/Developing_Operational_Requirements_Guides.pdf

CEN (2019). The 'New Approach'. https://boss.cen.eu/reference-material/guidancedoc/pages/newapproach/

CEN (n.d.a). About cen. cen communities. members. https://standards.cencenelec.eu/dyn/www/f?p=CEN:5

CEN (n.d.b). CEN/CLC/JTC 13 – Cybersecurity and Data Protection. Published standards. EN 17529:2022. https://standards.cencenelec.eu/dyn/www/f?p=CEN:110:0::::FSP_PROJECT,FSP_ORG_ID:63633,2307986&cs=11F702120AA40D5CC2DD42848140B1806

CEN (n.d.c). HAS assessment process (Innovative process). https://boss.cen.eu/developingdeliverables/pages/en/pages/has_assessment_process/

CEN, CENELEC, and ETSI (2021). Questions and Answers Regarding the European Standardization System. https://www.cencenelec.eu/media/CEN-CENELEC/European%20Standardization/Documents/IC/Cooperation%20Agreements/ansi-cen-cenelec-etsi_qas_ess_updated_september2021.pdf

CEN-CENELEC (2001). Guide 2 – Consumer interests and the preparation of European Standards. Edition 1. https://www.cencenelec.eu/media/Guides/CEN-CLC/cenclcguide2.pdf

CEN-CENELEC (2015a). CEN-CENELEC Guide 30. European Guide on Standards and Regulation - Better regulation through the use of voluntary standards - Guidance for policy makers. Edition 1. https://www.cencenelec.eu/media/Guides/CEN-CLC/cenclcguide30.pdf

CEN-CENELEC (2015b). Standardization Requests – Ad-hoc Groups (SRAHG), 'Their role, composition & way of working'. https://boss.cen.eu/media/iyheyfno/sreqahg.pdf

CEN-CENELEC (2017). Civil Society: Improving, strengthening and legitimising the European standardisation system. https://www.cencenelec.eu/media/CEN-CENELEC/Get%20Involved/Societal%20Stakeholders/civilsocietyleaflet.pdf

CEN-CENELEC (2018). Code of Conduct for delegates, experts and observers participating in the technical work of CEN and/or CENELEC. https://boss.cen.eu/reference-material/Guidancedoc/Pages/ExpertsCodeconduct

CEN-CENELEC (2020). CEN-CENELEC response to the EC White Paper on AI. https://www.cencenelec.eu/media/CEN-CENELEC/Areas%20of%20Work/Position%20Paper/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf

CEN-CENELEC (2020). Road Map on Artificial Intelligence (AI). Technical report, CEN-CENELEC Focus Group on AI. https://www.standict.eu/sites/default/files/2021-03/CEN-CLC_FGR_RoadMapAI.pdf

CEN-CENELEC (2021). Guide 25 – The concept of Cooperation with European Organizations and other stakeholders. Edition 3. https://www.cencenelec.eu/media/Guides/CEN-CLC/cenclcguide25.pdf

CEN-CENELEC (2024). CEN CENELEC in figures – 2024 Q3. https://www.cencenelec.eu/stats/CEN_CENELEC_in_figures_quarter.htm

CEN-CENELEC (2024a). Copyright protection of Harmonized Standards not in question – however, there is an overriding public interest in their disclosure. https://www.cencenelec.eu/news-and-events/news/2024/brief-news/2024-03-05-ecj-case/

CEN-CENELEC (2024b). Status Dashboards JTC21 European AI standardisation. https://www.linkedin.com/posts/sebastianhallensleben_status-dashboard-jtc21-european-ai-standardisation-activity-7235665875112988673-xfEH

CEN-CENELEC (2024c). Status of JTC 21 Activities in Response of the Standardization Request on AI. ETSI AI Conference 2024. https://docbox.etsi.org/Workshop/2024/02_ETSIAICONFERENCE/S03_STANDARDIZ_AIACT_LEGALFMK/JTC21ACTIVITIES_STD_REQUESTAI_CENCENELECJTC21.pdf

CEN-CENELEC (n.d.a). Artificial Intelligence. https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

CEN-CENELEC (n.d.b). CEN/CLC/JTC 21 Published Standards. https://standards.cencenelec.eu/dyn/www/f?p=205:32:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D

CEN-CENELEC (n.d.c). CEN/CLC/JTC 21 Subcommittees and Working Groups. https://standards.cencenelec.eu/dyn/www/f?p=205:29:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D#1

CEN-CENELEC (n.d.d). CEN/CLC/JTC 21 Work programme. https://standards.cencenelec.eu/dyn/www/f?p=205:22:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D

CENELEC (2017). Guide 13 – Frequently Asked Questions on the Frankfurt Agreement. Edition 1. https://www.cencenelec.eu/media/Guides/CLC/13_cenelecguide13_faq.pdf

CENELEC (n.d.). About cenelec. cenelec communities. members. https://standards.cencenelec.eu/dyn/www/f?p=CENELEC:5

CERNA (2014). Éthique de la recherche en robotique. Technical report, Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene. http://cerna-ethics-allistene.org/digitalAssets/38/38704_Avis_robotique_livret.pdf

Chan, K. (2023). Europe's world-leading artificial intelligence rules are facing a do-or-die moment. *Quartz*. https://qz.com/europes-world-leading-artificial-intelligence-rules-are-1851069721

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45. https://doi.org/10.1145/3641289

Charnovitz, S. (2005). International Standards and the WTO. https://dx.doi.org/10.2139/ssrn.694346

Chatila, R. and Havens, J. C. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version 2). Technical report, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Choi, B. H. (2024). NIST's software un-standards. *Lawfare*. https://www.lawfaremedia.org/article/nist's-software-un-standards

Chouldechova, A. (2017a). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163. Mary Ann Liebert, Inc. publishers. https://doi.org/10.1089/big.2016.0047

Chouldechova, A. (2017b). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163. Mary Ann Liebert, Inc. publishers. https://doi.org/10.1089/big.2016.0047

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2):2053951717718855. SAGE Publications. https://doi.org/10.1177/2053951717718855

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49(5-6):897–918. Springer. https://doi.org/10.1007/s11186-020-09411-3

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., and Baum, S. D. (2021). AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society*, 2(4):200–209. https://doi.org/10.1109/TTS.2021.3077595

CJEU (2012a). Judgment of the Court (Fourth Chamber), 12 July 2012. Fra.bo SpA v Deutsche Vereinigung des Gas- und Wasserfaches eV (DVGW) — Technisch-Wissenschaftlicher Verein. Case C-171/11. Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62011CJ0171

CJEU (2012b). Opinion of Advocate General Trstenjak delivered on 28 March 2012. Case C-171/11 P. Fra.bo SpA v Deutsche Vereinigung des Gas- und Wasserfaches eV (DVGW) — Technisch-Wissenschaftlicher Verein. Court of Justice of the European Union. https://curia.europa.eu/juris/document/document.jsf?text=&docid=121104&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1337476

CJEU (2014). Judgment of the Court (Grand Chamber), 22 January 2014. United Kingdom of Great Britain and Northern Ireland v European Parliament and Council of the European Union. Case C-270/12. Court of Justice of the European Union. https://curia.europa.eu/juris/document/document.jsf?docid=146621&doclang=EN

CJEU (2015). Judgment of the Court (Grand Chamber) of 6 October 2015 in Case C-362/14. Maximillian Schrems v Data Protection Commissioner. Request for a preliminary ruling from the High Court (Ireland). Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62014CJ0362

CJEU (2016a). Judgment of the Court (Third Chamber) of 27 October 2016. James Elliott Construction Limited v Irish Asphalt Limited. Case C-613/14. Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CA0613

CJEU (2016b). Opinion of Advocate General Campos Sanchez-Bordona delivered on 28 January 2016. Case C-613/14 P. James Elliott Construction Limited v Irish Asphalt Limited. Court of Justice of the European Union. https://curia.europa.eu/juris/document/document.jsf?docid=173893&doclang=en

CJEU (2020). Judgment of the Court (Grand Chamber) of 16 July 2020 in Case C-311/18. Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems. Request for a preliminary ruling from the High Court (Ireland). Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62018CJ0311

CJEU (2021). Judgment of the General Court (Fifth Chamber, Extended Composition) of 14 July 2021. Public.Resource.Org, Inc. and Right to Know CLG v European Commission. Case T-185/19. Court of Justice of the European Union. https://curia.europa.eu/juris/liste.jsf?language=en&td=ALL&num=T-185/19

CJEU (2023). Opinion of Advocate General Medina delivered on 22 June 2023. Case C-588/21 P. Public.Resource.Org, Inc., Right to Know CLG v European Commission. Court of Justice of the European Union. https://curia.europa.eu/juris/document/document.jsf?text=&docid=274881&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=2825624

CJEU (2024a). Judgment of the Court (Fourth Chamber) of 4 October 2024 in Case C-446/21. Maximilian Schrems v Meta Platforms Ireland Limited, anciennement Facebook Ireland Limited. Request for a preliminary ruling from the Oberster Gerichtshof. Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62021CJ0446

CJEU (2024b). Judgment of the Court (Grand Chamber) 5 March 2024 Public.Resource.Org and Right to Know v Commission and Others. Case C-588/21. Court of Justice of the European Union. https://curia.europa.eu/juris/document/document.jsf?text=&docid=283443&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=8223993

CNIL (2020). ISO 27701, an international standard addressing personal data protection. https://www.cnil.fr/en/iso-27701-international-standard-addressing-personal-data-protection

Collodi, C. (1881). Les aventures de Pinocchio. *La Bibliothèque électronique du Québec*, 886. Edition 2013. https://beq.ebooksgratuits.com/vents/Collodi-Pinocchio.pdf

Colmenarejo, A. B., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., and Kinder-Kurlanda, K. (2022). Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation. In *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 107–118. Association for Computing Machinery. https://doi.org/10.1145/3514094.3534158

Connelly, B. L., Certo, S. T., Ireland, R. D., and Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of management*, 37(1):39–67. SAGE Publications. https://doi.org/10.1177/0149206310388419

Council of Europe (2022). AI initiatives. https://www.coe.int/en/web/artificial-intelligence/national-initiatives

Council of Europe (2024a). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

Council of Europe (2024b). Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the point of view of Human Rights, Democracy and the Rule of Law (HUDERIA Methodology). https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights

Crawford, K. (2021). *Atlas of AI – Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://yalebooks.yale.edu/9780300264630/atlas-of-ai

Cuccuru, P. (2019). Interest Representation in European Standardisation: The Case of CEN and CENELEC. Amsterdam Law School Research Paper No. 2019-52, Amsterdam Centre for European Law and Governance Research Paper No. 2019-06. https://dx.doi.org/10.2139/ssrn.3505290

Cuccuru, P. (2020). Regulating by Request: On the Role and Status of the Standardisation Mandate under the New Approach. In Eliantonio, M. and Cauffman, C., editors, *The Legitimacy of Standardisation as a Regulatory Technique*, pages 48–63. https://doi.org/10.4337/9781789902952.00008

Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration (2023). Interim Measures for the Management of Generative Artificial Intelligence Services. *China Law Translate*. https://www.chinalawtranslate.com/generative-ai-interim/

Dal Bó, E. (2006). Regulatory Capture: A Review. *Oxford Review of Economic Policy*, 22(2):203–225. https://doi.org/10.1093/oxrep/grj013

Dartmouth (n.d.). Artificial Intelligence Coined at Dartmouth. https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

Dastin, J. (2018). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Datta, A. and Hartmann, T. (2025). Commission withdraws AI liability directive after Vance attack on regulation. *Euractiv*. https://www.euractiv.com/section/tech/news/commission-withdraws-ai-liability-directive-after-vance-attack-on-regulation/

David, P. A. (1986). Understanding the Economics of QWERTY: the Necessity of History. In Parker, W. N., editor, *Economic history and the modern economist*, pages 30–49. Basil Blackwell.

David, P. A. (1987). Some new standards for the economics of standardization in the information age. In Dasgupta, P. and Stoneman, P., editors, *Economic policy and technological performance*. Cambridge University Press. Centre for Economic Policy Research. https://doi.org/10.1017/CBO9780511559938

David, P. A. and Greenstein, S. (1990). The Economics Of Compatibility Standards: An Introduction To Recent Research. *Economics of Innovation and New Technology*, 1(1-2):3–41. Routledge. https://doi.org/10.1080/10438599000000002

de Almeida, P. G. R., dos Santos, C. D., and Farias, J. S. (2021). Artificial Intelligence Regulation: a framework for governance. *Ethics and Information Technology*, 23(3):505–525. https://doi.org/10.1007/s10676-021-09593-z

De Gregorio, G. (2021). The rise of digital constitutionalism in the European Union. *International Journal of Constitutional Law*, 19(1):41–70. https://doi.org/10.1093/icon/moab001

de Tervueren, A. (2012). ANEC Position Paper on CE marking "Caveat Emptor - Buyer Beware". Technical report, The European Association for the Co-ordination of Consumer Representation in Standardisation (ANEC). https://anec.eu/publications/position-papers/201-anec-position-paper-on-ce-marking-caveat-emptor-buyer-beware de Vries, H., Blind, K., Mangelsdorf, A., Verheul, H., and van der Zwan, J. (2009). SME access to European standardization. Rotterdam School of Management, Erasmus University. https://www.erim.eur.nl/fileadmin/default/content/erim/content_area/news/2009/smeaccessreport%202009.pdf

de Vries, H. J. (1999). *Standardization: A Business Approach to the Role of National Standardization Organizations*. Springer New York. https://doi.org/10.1007/978-1-4757-3042-5

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., and Kühl, N. (2024). Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. *arXiv preprint arXiv:2403.20089*. http://arxiv.org/abs/2403.20089

DeepMind (n.d.a). AlphaFold. https://deepmind.google/technologies/alphafold/

DeepMind (n.d.b). AlphaGo. https://deepmind.google/research/breakthroughs/alphago/

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://aclanthology.org/N19-1423

Di Maio, P. (2018). Toward A Web Standard for Explainable AI? W3C AI KR (Artificial Intelligence Knowledge Representation) Community Group. https://www.w3.org/community/aikr/2018/10/31/towards-a-web-standard-for-explainable-ai/

Dignum, V. (2024). How Europe is Shaping AI for Human Rights. *AI Policy Lab*. https://aipolicylab.se/2024/09/05/how-europe-is-shaping-ai-for-human-rights/

DIN and DKE (2021). Joint position paper on the European Commission's proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts – COM(2021) 206. https://www.din.de/resource/blob/800324/c50ed443e81c47f8860b3f5c2b3b0742/21-06-din-dke-position-paper-artificial-intelligence-act-data.pdf

DIN and DKE (2022). German Standardization Roadmap on Artificial Intelligence (2nd Edition). http://www.din.de/go/roadmap-ai

DTI (2005). The Empirical Economics of Standards. Technical report. http://www.dti.gov.uk/iese/The-Empirical-Economics-of-Standards.pdf

du Net, L. Q. (2024). With the AI Act adopted, the techno-solutionist gold-rush can continue. https://www.laquadrature.net/en/2024/05/22/with-the-ai-act-adopted-the-techno-solutionist-gold-rush-can-continue/

Dumitru, F. (2008). The Organisms of European Standardizing. *The Annals of the University of Oradea. Economic Sciences*, 1(55):300–304. International Business and European Intergration. http://anale.steconomiceuoradea.ro/en/2010/05/05/tom-2008-volumul-i-intenational-business/

Dunkerley, D. and Fudge, S. (2004). The role of civil society in European integration A framework for analysis. *European Societies*, 6(2):237–254. Routledge. https://doi.org/10.1080/1461669041000169067З

Durmus, M. (2021). Overview of National AI-Strategies. https://www.aisoma.de/wp-content/uploads/2021/01/Overview-of-National-AI-Strategies.pdf

E-Santé (2023). Lancement du premier label de Garantie Humaine de l'Intelligence artificielle. Evenement de l'Agence du Numérique en Santé. https://esante.gouv.fr/agenda/lancement-du-premier-label-de-garantie-humaine-de-lintelligence-artificielle

Ebers, M. (2022). Standardizing AI - The Case of the European Commission's Proposal for an Artificial Intelligence Act. In DiMatteo, L. A., Poncibò, C., and Cannarsa, M., editors, *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*. Cambridge University Press. https://doi.org/10.1017/9781009072168.030

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., and Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act – A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, 4(4):589–603. https://doi.org/10.3390/j4040043

ECJ (1958a). Judgment of the Court of 13 June 1958. Meroni & Co, Industrie Metallurgiche, società in accomandita semplice v High Authority of the European Coal and Steel Community. Case 10-56. European Court of Justice. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61956CJ0010

ECJ (1958b). Judgment of the Court of 13 June 1958. Meroni & Co, Industrie Metallurgiche, SpA v High Authority of the European Coal and Steel Community. Case 9-56. European Court of Justice. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61956CJ0009

EDPB (2022a). EDPB adopts "wish list"' of procedural aspects, first EU data protection seal and a statement on digital euro. European Data Protection Board. https://www.edpb.europa.eu/news/news/2022/edpb-adopts-wish-list-procedural-aspects-first-eu-data-protection-seal-and-statement_en

EDPB (2022b). Opinion 28/2022 on the Europrivacy criteria of certification regarding their approval by the Board as European Data Protection Seal pursuant to Article 42.5 (GDPR). European Data Protection Board. https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282022-europrivacy-criteria-certification_en

EDRi (2022). The role of standards and standardisation processes in the EU's Artificial Intelligence (AI) Act. *European Digital Rights (EDRi)*. https://edri.org/wp-content/uploads/2022/05/The-role-of-standards-and-standardisation-processes-in-the-EUs-Artificial-Intelligence-AI-Act.pdf

EDRi (2024). EU's AI Act fails to set gold standard for human rights. *European Digital Rights (EDRi)*. https://edri.org/our-work/eu-ai-act-fails-to-set-gold-standard-for-human-rights/

Edwards, L. (2022). Expert opinion. Regulating AI in Europe: four problems and four solutions. Technical report, Ada Lovelace Institute. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf

EESC (2017). Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society' (own-initiative opinion). Technical report, European Economic and Social Committee (EESC). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016IE5369

Egan, M. (2001). *Constructing a European Market: Standards, Regulation, and Governance*. Oxford University Press. https://doi.org/10.1093/0199244057.001.0001

Eliantonio, M. (2017). Judicial Control of the EU Harmonized Standards: Entering a Black Hole? *Legal Issues of Economic Integration*, 44(4):395–407. https://doi.org/10.54648/leie2017022

Eliantonio, M. and Cauffman, C. (2020). The Legitimacy of Standardisation as a Regulatory Technique in the EU – A Cross-disciplinary and Multi-level Analysis: An Introduction. In Eliantonio, M. and Cauffman, C., editors, *The Legitimacy of Standardisation as a Regulatory Technique*, Law 2020, pages 1–18. Edward Elgar Publishing. https://doi.org/10.4337/9781789902952.00005

Eliantonio, M. and Medzmariashvili, M. (2017). Hybridity Under Scrutiny: How European Standardization Shakes the Foundations of EU Constitutional and Internal Market Law. *Legal Issues of Economic Integration*, 44(4). https://doi.org/10.54648/leie2017017

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR. https://proceedings.mlr.press/v81/ensign18a.html

Ernst & Young (2020). Study on the implementation of the Regulation (EU) No. 1025/2012 (Article 24): final report. Technical report, European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. https://data.europa.eu/doi/10.2873/504681

EthicalML (n.d.). GitHub - Awesome AI Guidelines, The Institute for Ethical Machine Learning. https://github.com/EthicalML/awesome-artificial-intelligence-guidelines

ETSI (n.d.). Membership of ETSI: ETSI members around the world. https://www.etsi.org/membership

ETUC (2021). Feedback on the (roadmap) consultation of citizens and stakeholders on the forthcoming "EU Standardisation strategy". European Trade Union Confederation (ETUC). https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13099-Standardisation-strategy/F2663296_en

ETUC (2024). Artificial Intelligence standardisation Inclusiveness Newsletter. Edition 1. European Trade Union Confederation. https://www.etuc.org/sites/default/files/page/file/2024-03/AI%20standardisation%20Inclusiveness_Newsletter1.pdf

European Commission (1990). Commission Green Paper on the Development of European Standardization: Action for Faster Technological Integration in Europe. COM/90/456 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:51990DC0456

European Commission (2001). European governance - A white paper. OJ C 287, p.1–29. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52001DC0428

European Commission (2011a). Commission Decision of 28 November 2011 setting up the European multi-stakeholder platform on ICT standardisation. OJ C 349, p.4–6. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32011D1130(02)

European Commission (2011b). Commission Staff Working Paper – Impact Assessment Accompanying document to the Proposal for a Regulation of the European Parliament and of the Council on European Standardisation and amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/105/EC and 2009/23/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011SC0671

European Commission (2011c). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee – A strategic vision for European standards: Moving forward to enhance and accelerate the sustainable growth of the European economy by 2020. COM/2011/0311 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011DC0311

European Commission (2012a). Charter of Fundamental Rights of the European Union. OJ C 326, p.391–407. http://data.europa.eu/eli/treaty/char_2012/oj/eng

European Commission (2012b). Consolidated version of the Treaty on European Union. OJ C326, p.13–390. http://data.europa.eu/eli/treaty/teu_2012/oj

European Commission (2012c). Consolidated version of the Treaty on the Functioning of the European Union. OJ C 326, p.47–390. http://data.europa.eu/eli/treaty/tfeu_2012/oj/eng

European Commission (2015). Vademecum on European standardisation in support of Union legislation and policies - Part 1 Role of the Commission's Standardisation requests to the European standardisation organisations. (Working Document SWD(2015) 205 final PART 1/3). https://ec.europa.eu/docsroom/documents/13507/attachments/1/translations

European Commission (2016a). Commission Implementing Decision (EU) 2016/537 of 5 April 2016 on the publication with a restriction in the Official Journal of the European Union of the reference of standard EN 50566:2013 on requirements to demonstrate compliance of radio frequency fields from handheld and body-mounted wireless communication devices used by the general public (30 MHz-6 GHz) under Directive 1999/5/EC of the European Parliament and of the Council. OJ L, P.17-19. http://data.europa.eu/eli/dec_impl/2016/537/oj

European Commission (2016b). Commission Staff Working Document on the implemetation of the actions foreseen in the 2015 and 2016 Union work programmes for European standardisation, including the implementing acts and mandates sent to the European standardisation organisations. Accompanying the document Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee. The Annual Union work programme for European standardisation for 2016. SWD/2015/0301 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52015SC0301

European Commission (2018a). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Artificial Intelligence For Europe. COM/2018/237 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN

European Commission (2018b). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Coordinated Plan on Artificial Intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0795

European Commission (2019). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – Building Trust in Human Centric Artificial Intelligence. COM(2019)168. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168

European Commission (2020a). Commission Implementing Regulation (EU) 2020/1207 of 19 August 2020 laying down rules for the application of Regulation (EU) 2017/745 of the European Parliament and of the Council as regards common specifications for the reprocessing of single-use devices. OJ L 273, p.3–15. http://data.europa.eu/eli/reg_impl/2020/1207/oj

European Commission (2020b). ICT Standardisation Observatory and Support Facility in Europe. CORDIS - EU research results. https://doi.org/10.3030/951972

European Commission (2020c). Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee – Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. COM/2020/64 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0064

European Commission (2020d). White Paper on Artificial Intelligence: a European approach to excellence and trust - European Commission. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission (2021a). Artificial intelligence – ethical and legal requirements. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/feedback_en?p_id=24212003

European Commission (2021b). Commission Staff Working Document – Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. SWD/2021/84 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021SC0084

European Commission (2021c). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – Fostering a European approach to Artificial Intelligence. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2021:205:FIN

European Commission (2021d). Coordinated Plan on Artificial Intelligence 2021 Review. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

European Commission (2021e). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union legislative acts. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission (2022a). Commission Decision setting up the group of experts 'High-Level Forum on European Standardisation". C(2022)6189. https://ec.europa.eu/transparency/documents-register/detail?ref=C(2022)6189&lang=en

European Commission (2022b). Commission Implementing Regulation (EU) 2022/1107 of 4 July 2022 laying down common specifications for certain class d in vitro diagnostic medical devices in accordance with Regulation (EU) 2017/746 of the European Parliament and of the Council. OJ L 178, p.3–56. http://data.europa.eu/eli/reg_impl/2022/1107/oj

European Commission (2022c). Commission Implementing Regulation (eu) 2022/2346 of 1 December 2022 laying down common specifications for the groups of products without an intended medical purpose listed in Annex XVI to Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices. OJ L 311, p.60–93. http://data.europa.eu/eli/reg_impl/2022/2346/oj

European Commission (2022d). Commission Notice - The 'Blue Guide' on the implementation of EU product rules 2022 (OJ C 247; pp. 1–152). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022XC0629(04)

European Commission (2022e). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – An EU Strategy on Standardisation Setting global standards in support of a resilient, green and digital EU single market (COM/2022/31 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022DC0031

European Commission (2022f). Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence.

European Commission (2022g). New approach to enable global leadership of EU standards promoting values and a resilient, green and digital Single Market. Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/news/new-approach-enable-global-leadership-eu-standards-promoting-values-and-resilient-green-and-digital-2022-02-02_en

European Commission (2023a). Commission implementing decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence (C(2023)3215). https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en

European Commission (2023b). Commission Recommendation (EU) 2023/498 of 1 March 2023 on a Code of Practice on standardisation in the European Research Area. OJ L 69, p. 63-74. http://data.europa.eu/eli/reco/2023/498/oj/eng

European Commission (2023c). EUropean ACTive stakeholders in AI standardization. EU Funding Tenders Portal. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-det ails/43252476/101140954/SMP?order=DESC&pageNumber=1&pageSize=50&sortBy=title&keywords=euact&isExactMatch=false

European Commission (2023d). European Standards for Artificial Intelligence. EU Funding Tenders Portal. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/43252476/101112776/SMP?order=DESC&pageNumber=1&pageSize=50&sortBy=title&keywords=esai&isExactMatch=false

European Commission (2024a). 1st European AI Office webinar on Risk management logic of the AI Act and related standards. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/events/1st-european-ai-office-webinar-risk-management-logic-ai-act-and-related-standards

European Commission (2024b). AI Act: Have Your Say on Trustworthy General-Purpose AI. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/consultations/ai-act-have-your-say-trustworthy-general-purpose-ai

European Commission (2024c). Commission Decision of 24 January 2024 establishing the European Artificial Intelligence Office. OJ C, C/2024/1459. http://data.europa.eu/eli/C/2024/1459/oj

European Commission (2024d). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on boosting startups and innovation in trustworthy artificial intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2024:28:FIN

European Commission (2024e). The kick-off Plenary for the General-Purpose AI Code of Practice took place online. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/kick-plenary-general-purpose-ai-code-practice-took-place-online

European Commission (2024f). Meet the Chairs leading the development of the first General-Purpose AI Code of Practice. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/meet-chairs-leading-development-first-general-purpose-ai-code-practice

European Commission (n.d.a). AI Act. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission (n.d.b). AI Pact. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/ai-pact

European Commission (n.d.c). CE marking. *Your Europe*. https://europa.eu/youreurope/business/product-requirements/labels-markings/ce-marking/index_en.htm

European Commission (n.d.d). Ecodesign and Energy Labelling. Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards/ecodesign_en

European Commission (n.d.e). Ecodesign for Sustainable Products Regulation. https://commission.europa.eu/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/ecodesign-sustainable-products-regulation_en

European Commission (n.d.f). European AI Office. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/ai-office

European Commission (n.d.g). European Multi-Stakeholder Platform on ICT Standardisation. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/multi-stakeholder-platform-ict-standardisation

European Commission (n.d.h). European standardisation – evaluation. Public Consultations and Feedback. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13446-European-standardisation-evaluation_en

European Commission (n.d.i). Excellence and trust in artificial intelligence. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_en

European Commission (n.d.j). Grants. https://international-partnerships.ec.europa.eu/funding-and-technical-assistance/looking-funding/grants_en

European Commission (n.d.k). Harmonised Standards. *Internal Market, Industry, Entrepreneurship and SMEs*. https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en

European Commission (n.d.l). Harmonised standards. Public Health. https://health.ec.europa.eu/medical-devices-topics-interest/harmonised-standards_en

European Commission (n.d.m). High-Level Forum on European Standardisation. Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/single-market/european-standards/standardisation-policy/high-level-forum-european-standardisation_en

European Commission (n.d.n). Horizon 2020. Research and innovation. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

European Commission (n.d.o). Horizon Europe. Research and innovation. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en

European Commission (n.d.p). Key players in European Standardisation. *Internal Market, Industry, Entrepreneurship and SMEs*. https://single-market-economy.ec.europa.eu/single-market/european-standards/key-players-european-standardisation_en

European Commission (n.d.q). Manufacturers. https://single-market-economy.ec.europa.eu/single-market/ce-marking/manufacturers_en

European Commission (n.d.r). New legislative framework. *Internal Market, Industry, Entrepreneurship and SMEs*. https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

European Commission (n.d.s). Notified bodies. https://single-market-economy.ec.europa.eu/single-market/goods/building-blocks/notified-bodies_en

European Commission (n.d.t). Notified bodies (NANDO). *Single Market Compliance Space*. https://webgate.ec.europa.eu/single-market-compliance-space/notified-bodies

European Commission (n.d.u). Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/activities/testing-and-experimentation-facilities

European Commission (n.d.v). Standardisation at the Joint Research Centre. EU Science Hub. https://joint-research-centre.ec.europa.eu/tools-and-laboratories/standardisation_en

European Commission DG SANTE (n.d.). Medical Device Coordination Group Working Groups. https://health.ec.europa.eu/medical-devices-dialogue-between-interested-parties/medical-device-coordination-group-working-groups_en

European Council (1985a). Consolidated text: Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. http://data.europa.eu/eli/dir/1985/374/1999-06-04

European Council (1985b). Council Resolution of 7 May 1985 on a new approach to technical harmonization and standards. OJ C 136, p.1–9. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31985Y0604(01)

European Council (1989). Council Resolution of 21 December 1989 on a global approach to conformity assessment. OJ C 10, p.1–2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31990Y0116(01)

European Council (1993). Council Decision of 22 July 1993 concerning the modules for the various phases of the conformity assessment procedures and the rules for the affixing and use of the CE conformity marking, which are intended to be used in the technical harmonization directives. 93/465/EEC. OJ L 220, p. 23–39 (repealed). http://data.europa.eu/eli/dec/1993/465/oj/eng

European Council (2020). Special meeting of the European Council (1 and 2 October 2020) – Conclusions. https://www.consilium.europa.eu/en/press/press-releases/2020/10/02/european-council-conclusions-1-2-october-2020/

European Council (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Presidency compromise text. 14278/21. https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

European Council (2022a). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and Amending certain Union legislative acts - General approach. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

European Council (2022b). Proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union - Text de compromis de la présidence - Article 3, paragraphe 1 ter, Articles 4 bis à 4 quater, Annexe VI (3) et (4), considérant 12 bis bis. https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-FRA-Art-34-13-May.pdf

European Council (2022c). Proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union - Text de compromis de la présidence - Version consolidée. https://data.consilium.europa.eu/doc/document/ST-10069-2022-INIT/x/pdf

European Council (2024). Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

European Council of Engineers Chambers (2021). Statement on "Roadmap Standardization Strategy". European Commission. Consultation and Feedback: Standardisation startegy. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13099-Standardisation-strategy/F2665555_en

European Parliament (2010). European Parliament resolution of 21 October 2010 on the future of European standardisation (2010/2051(INI)). https://www.europarl.europa.eu/doceo/document/TA-7-2010-0384_EN.html

European Parliament (2017). European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

European Parliament (2020a). The ethics of artificial intelligence: issues and initiatives. Technical report, Directorate General for Parliamentary Research Services. Publications Office of the European Union. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452

European Parliament (2020b). European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)). https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html

European Parliament (2020c). European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)). https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.html

European Parliament (2022). Long-awaited common charger for mobile devices will be a reality in 2024. News European Parliament. https://www.europarl.europa.eu/news/en/press-room/20220930IPR41928/long-awaited-common-charger-for-mobile-devices-will-be-a-reality-in-2024

European Parliament (2023a). Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). P9_TA(2023)0236. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

European Parliament (2023b). Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

European Parliament (2023c). Report on a standardisation strategy for the single market. A9-0136/2023. Committee on the Internal Market and Consumer Protection. Rapporteur: Bielan, A. https://www.europarl.europa.eu/doceo/document/A-9-2023-0136_EN.html

European Parliament (2024). Artificial Intelligence Act: MEPs adopt landmark law. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

European Parliament (n.d.a). Framework of ethical aspects of artificial intelligence, robotics and related technologies. *Legislative Train Schedule*. https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-ethical-framework

European Parliament (n.d.b). List of legal bases providing for the ordinary legislative procedure in the Treaty of Lisbon. https://www.europarl.europa.eu/cmsdata/198025/List_of_legal_bases.pdf

European Parliament and Council (1999). Directive 1999/5/EC of the European Parliament and of the Council of 9 March 1999 on radio equipment and telecommunications terminal equipment and the mutual recognition of their conformity. OJ L, p.10-28. Repealed. https://eur-lex.europa.eu/eli/dir/1999/5/oj

European Parliament and Council (2008a). Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC. OJ L 218, p.82–128. http://data.europa.eu/eli/dec/2008/768(1)/oj/eng

European Parliament and Council (2008b). Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93. OJ L 218, p.30—47. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex:32008R0765

European Parliament and Council (2009a). Directive 2009/125/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for the setting of ecodesign requirements for energy-related products (recast). OJ L 285, p.10-35. http://data.europa.eu/eli/dir/2009/125/oj/eng

European Parliament and Council (2009b). Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys. OJ L 170, p.1-37. http://data.europa.eu/eli/dir/2009/48/oj/eng

European Parliament and Council (2012). Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/eec and 93/15/eec and Directives 94/9/ec, 94/25/ec, 95/16/ec, 97/23/ec, 98/34/ec, 2004/22/ec, 2007/23/ec, 2009/23/ec and 2009/105/ec of the European Parliament and of the Council and repealing Council Decision 87/95/eec and Decision no 1673/2006/ec of the European Parliament and of the Council. OJ L 316, p.12–33. https://eur-lex.europa.eu/eli/reg/2012/1025/oj

European Parliament and Council (2014). Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC. https://eur-lex.europa.eu/eli/dir/2014/53/oj

European Parliament and Council (2016a). Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA. OJ L 119, p.89-131. https://eur-lex.europa.eu/eli/dir/2016/680/2016-05-04

European Parliament and Council (2016b). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 18, p.1–88. http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng

European Parliament and Council (2017a). Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications). COM/2017/010 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017PC0010

European Parliament and Council (2017b). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. OJ L 117, p.1-175. http://data.europa.eu/eli/reg/2017/745/oj/eng

European Parliament and Council (2018). Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC. OJ L 295, p.39-98. https://eur-lex.europa.eu/eli/reg/2018/1725/oj

European Parliament and Council (2019). Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011. OJ L 169, p.1–44. http://data.europa.eu/eli/reg/2019/1020/oj

European Parliament and Council (2020). Publication of an update to the list of national standardisation bodies pursuant to Article 27 of Regulation (EU) No 1025/2012 of the European Parliament and of the Council on European standardisation 2020/C 104/03. PUB/2020/249. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020XC0331(01)

European Parliament and Council (2022). Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0496

European Parliament and Council (2022). Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). http://data.europa.eu/eli/reg/2022/1925/oj/eng

European Parliament and Council (2022a). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). OJ L 277, p.1–102. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

European Parliament and Council (2022b). Regulation (EU) 2022/2480 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 1025/2012 as regards decisions of European standardisation organisations concerning European standards and European standardisation deliverables. OJ L 323, p.1-3. http://data.europa.eu/eli/reg/2022/2480/oj

European Parliament and Council (2022c). Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). OJ L 152, p.1–44. http://data.europa.eu/eli/reg/2022/868/oj/eng

European Parliament and Council (2023a). Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC. OJ L 165, p.1-102. https://eur-lex.europa.eu/eli/reg/2023/1230/oj

European Parliament and Council (2023b). Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). OJ L. https://eur-lex.europa.eu/eli/reg/2023/2854

European Parliament and Council (2023c). Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC. OJ L 135, p.1–51. http://data.europa.eu/eli/reg/2023/988/oj

European Parliament and Council (2024a). Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (Text with EEA relevance). http://data.europa.eu/eli/dir/2024/2853/oj

European Parliament and Council (2024b). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ L. http://data.europa.eu/eli/reg/2024/1689/oj

European Parliament and Council (2024c). Regulation (EU) 2024/1781 of the European Parliament and of the Council of 13 June 2024 establishing a framework for the setting of ecodesign requirements for sustainable products, amending Directive (EU) 2020/1828 and Regulation (EU) 2023/1542 and repealing Directive 2009/125/EC. Legislative Body: EP, CONSIL. http://data.europa.eu/eli/reg/2024/1781/oj/eng

European Parliament and Council (2024d). Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act). OJ L. http://data.europa.eu/eli/reg/2024/2847/oj

European Parliament, Council and Commission (2003). Interinstitutional agreement on better law-making (OJ C 321, p.1-5). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32003Q1231(01)

European Union (2011). A new approach to technical harmonisation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l21001a

Europrivacy (n.d.a). Audit and Certification in Data Protection. https://www.europrivacy.org/

Europrivacy (n.d.b). Europrivacy Benefits and Advantages. https://www.europrivacy.org/en/ep/benefits

Evers, C. (2024). Talking past each other? Navigating discourse on ethical AI: Comparing the discourse on ethical AI policy by Big Tech companies and the European Commission. In *FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1885–1896. Association for Computing Machinery. https://doi.org/10.1145/3630106.3659013

Everson, M., Majone, G., Metcalfe, L., and Schout, A. (1999). The Role of Specialised Agencies in Decentralising EU Governance. Technical report, Report Presented to the Commission. https://www.academia.edu/103116487/The_Role_of_Specialised_Agencies_in_Decentralising_EU_Governance

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint arXiv:1707.08945*. http://arxiv.org/abs/1707.08945

Farrell, J. and Saloner, G. (1985). Economic issues in standardization. Massachusetts Institute of Technology. https://archive.org/details/economicissuesin00farr2

Ferrari, E. (2020). Technocracy Meets Populism: The Dominant Technological Imaginary of Silicon Valley. *Communication, Culture and Critique*, 13(1):121–124. https://doi.org/10.1093/ccc/tcz051

Fischhoff, B. (1983). "Acceptable Risk": The Case of Nuclear Power. *Journal of Policy Analysis and Management*, 2(4):559–575. Wiley, Association for Public Policy Analysis and Management. https://doi.org/10.2307/3323574

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*. https://dx.doi.org/10.2139/ssrn.3518482

Floridi, L. and Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

for Standardization e. V., D. G. I. (2000). *Economic benefits of standardization – Summary of results*. Beuth Verlag. https://doi.org/10.2366/34120260

Fraser, H. and Bello y Villarino, J.-M. (2023). Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough. *European Journal of Risk Regulation*, pages 1–16. https://doi.org/10.1017/err.2023.57

Frattone, C. (2022). Reasonable AI and other creatures. What role for AI standards in liability litigation? *Journal of Law, Market & Innovation*, 1(3):15–55. https://doi.org/10.13135/2785-7867/7166

Freiberg, A. (2010). *The Tools of Regulation*. Federation Press. https://research.monash.edu/en/publications/the-tools-of-regulation

Fung, P. and Etienne, H. (2022). Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU. *AI and Ethics*, 3. Springer. https://doi.org/10.1007/s43681-022-00180-6

Future of Life Institute (2017). Asilomar AI Principles. https://futureoflife.org/open-letter/ai-principles

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Future of Life Institute (n.d.). Historic Timeline – EU Artificial Intelligence Act. https://artificialintelligenceact.eu/developments/

Galvagna, C. (2023). Inclusive AI governance. Technical report, Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/

Ganascia, J.-G. (2017). *Le Mythe de la Singularité - Faut-il craindre l'intelligence artificielle ?* Seuil. https://www.seuil.com/ouvrage/le-mythe-de-la-singularite-jean-gabriel-ganascia/9782021309997

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92. https://doi.org/10.1145/3458723

Gebru, T. and Torres, P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4). https://firstmonday.org/ojs/index.php/fm/article/view/13636

Gellert, R. (2018). Understanding the notion of risk in the General Data Protection Regulation. *Computer Law & Security Review*, 34(2):279–288. https://www.doi.org/10.1016/j.clsr.2017.12.003

Gestel, R. v. and Micklitz, H.-W. (2013). European integration through standardization: How judicial review is breaking down the club house of private standardization bodies. *Common Market Law Review*, 50(1):141–181. https://doi.org/10.54648/cola2013007

Giovannini, C. (2021). The role of standards in meeting consumer needs and expectations of AI in the European Commission proposal for an Artificial Intelligence Act. Technical report, The European consumer voice in standardisation (ANEC). https://www.anec.eu/images/Publications/position-papers/Digital/ANEC-DIGITAL-2021-G-141.pdf

Gkritsi, E. (2024). Commission's DG for technology restructures to realise AI Office. *www.euractiv.com*. https://www.euractiv.com/section/artificial-intelligence/news/commissions-dg-for-technology-restructures-to-realise-ai-office/

Gnes, M. (2017). Do Administrative Law Principles Apply to European Standardization: Agencification or Privatization? *Legal Issues of Economic Integration*, 44(4). https://doi.org/10.54648/leie2017020

Goertzel, B. and Pennachin, C. (2007). *Artificial General Intelligence*. Springer. https://doi.org/10.1007/978-3-540-68677-4

Goffi, E. R. and Momcilovic, A. (2022). Respecting cultural diversity in ethics applied to AI : a new approach for a multicultural governance. *Misión Jurídica*, 15(23):111–122. https://doi.org/10.25058/1794600X.2135

Golpayegani, D., Pandit, H. J., and Lewis, D. (2023). To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *FAccT'23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 905–915. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594050

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Workshop on Widening NLP*. https://aclanthology.org/W19-3621

Gonzalez Torres, A. P. and Ali-Vehmas, T. (2024). Governing Through Standards: Artificial Intelligence and Values. Delft University of Technology. https://easychair.org/publications/preprint/xgpk

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. 6:31–88. Elsevier. https://doi.org/10.1016/S0065-2458(08)60418-0

GoodAlgo (n.d.). Labels éthiques et engagés. https://goodalgo.fr/labels-ethiquement-engages/

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.48550/arXiv.1412.6572

Google (2021). Consultation on the EU AI Act Proposal: Google's submission. European Commission - Have your say. Feedback from: Google. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

Gornet, M. (2024a). The AI Act: the evolution of "trustworthy AI"' from policy documents to mandatory regulation. https://hal.science/hal-04785519

Gornet, M. (2024b). La stratégie européenne de régulation de l'intelligence artificielle : entre normes techniques et droits fondamentaux. Série de blog "Voix émergentes dans l'IA et la société", University of Ottawa. https://www.uottawa.ca/recherche-innovation/toutes-nouvelles/strategie-europeenne-regulation-lia-entre-normes-techniques-droits-fondamentaux

Gornet, M. (2024c). Regulation through standardisation: the legitimacy issue of european AI standards. https://hal.science/hal-04785178

Gornet, M. (2024d). Too broad to handle: can we "fix" harmonised standards on artificial intelligence by focusing on vertical sectors? https://hal.science/hal-04785208

Gornet, M., Delarue, S., Boritchev, M., and Viard, T. (2024). Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 127–140, Rio de Janeiro, Brazil. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658545

Gornet, M. and Herman, H. (2024). A peek into european standards making for ai: between geopolitical and economic interests. https://hal.science/hal-04784035

Gornet, M., Kirchner, C., and Tessier, C. (2022a). Operational Fairness for Facial Authentication Systems. *ERCIM News N°131*, Special theme on Ethical Software Engineering and Ethically Aligned Design. https://ercim-news.ercim.eu/en131/special/operational-fairness-for-facial-authentication-systems

Gornet, M., Kirchner, C., and Tessier, C. (2022b). Operational fairness when coding facial authentication. https://hal.science/hal-04447868

Gornet, M. and Maxwell, W. (2022). Intelligence artificielle: normes techniques et droits fondamentaux, un mélange risqué. *The Conversation*. https://theconversation.com/intelligence-artificielle-normes-techniques-et-droits-fondamentaux-un-melange-risque-189587

Gornet, M. and Maxwell, W. (2023a). L'IA explicable appliquée à la détection de ceintures et de téléphones au volant. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA @ PfIA 2023)*, pages 46–56, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). https://hal.science/hal-04158889

Gornet, M. and Maxwell, W. (2023b). Normes techniques et éthique de l'IA. In *Conférence Nationale en Intelligence Artificielle (CNIA @ PfIA 2023)*, Strasbourg, France. Association Française pour l'Intelligence Artificielle (AFIA). https://pfia23.icube.unistra.fr/conferences/cnia/publications/Gornet.pdf

Gornet, M. and Maxwell, W. (2023c). Recherche sur les normes techniques pour l'IA et les droits fondamentaux à Télécom Paris. Bulletin N°120 "IA & Normes". pages 14–19. Association française pour l'Intelligence Artificielle (AfIA). https://www.hub-franceia.fr/wp-content/uploads/2024/03/AFIA-bulletin-120_avr23.pdf

Gornet, M. and Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, 13(3). https://doi.org/10.14763/2024.3.1784

Gornet, M. and Viard, T. (2023a). Mapping AI Ethics: A Quantitative Analysis of the Plurality, and Lack Thereof, of Discourses. https://dx.doi.org/10.2139/ssrn.4582657

Gornet, M. and Viard, T. (2023b). Queer Identities and Machine Learning. https://hal.science/hal-04763352

Government of Canada (2022). Artificial Intelligence and Data Act. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

Grafenstein, M. v. (2022). Co-regulation and competitive advantage in the GDPR: Data protection certification mechanisms, codes of conduct and data protection-by-design. In González Fuster, G., van Brakel, R., and De Hert, P., editors, *Research Handbook on Privacy and Data Protection Law*, Law 2022, pages 402–432. Edward Elgar Publishing. https://doi.org/10.4337/9781786438515

Graz, J.-C. (2006). International Standardisation and Corporate Democracy. In Giesen, K.-G. and van der Pijl, K., editors, *Global Norms in the Twenty-First Century*, pages 118–133. Cambridge Scholars Press. https://doi.org/10.7202/019309ar

Graz, J.-C. (2019). *The Power of Standards: Hybrid Authority and the Globalisation of Services*. Cambridge University Press. https://doi.org/10.1017/9781108759038

Grindley, P. (1995). *Standards, Strategy, and Policy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198288077.001.0001

Grother, P. (2022). Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials. Technical Report NIST IR 8429, National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.IR.8429.ipd

Grother, P. J., Ngan, M. L., and Hanaoka, K. K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Technical Report NIST IR 8280, National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.IR.8280

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., and Strait, A. (2024). Auditing Work: Exploring the New York City algorithmic bias audit regime. In *FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1107–1120. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658959

Gupta, M., Akiri, C., Aryal, K., Parker, E., and Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, 11:80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Gwilliam, M., Hegde, S., Tinubu, L., and Hanson, A. (2021). Rethinking Common Assumptions to Mitigate Racial Bias in Face Recognition Datasets. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4106–4115. https://ieeexplore.ieee.org/document/9607499

Haataja, M. and Bryson, J. J. (2021). What costs should we expect from the EU's AI Act? https://osf.io/8nzb4

Hacker, P. (2023). What's Missing from the EU AI Act. *Verfassungsblog*. Verfassungsblog. https://www.doi.org/10.59704/3f4921d4a3fbeeee

Hacker, P. (2024). Comments on the Final Trilogue Version of the AI Act. https://dx.doi.org/10.2139/ssrn.4757603

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, 30(1):99–120. Springer. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, 2(4):851–867. Springer. https://doi.org/10.1007/s43681-021-00122-8

Haimowitz, J. and Warren, J. (2007). Economic value of standardization. Technical report, Standards Council of Canada by The Conference Board of Canada. https://publications.gc.ca/collections/collection_2017/ccn-scc/Iu81-3-9-2007-eng.pdf

Hanson, D. (2005). *CE Marking, Product Standards and World Trade*. Edward Elgar Publishing. https://www.e-elgar.com/shop/gbp/ce-marking-product-standards-and-world-trade-9781843767732.html

Harding, S. (1992). After the Neutrality Ideal: Science, Politics, and "Strong Objectivity". *Social Research*, 59(3):567–587. The Johns Hopkins University Press. https://www.jstor.org/stable/40970706

Heading, S. and Cavaciuti-Wishart, E. (2024). These are the biggest global risks we face in 2024 and beyond. *World Economic Forum*. https://www.weforum.org/stories/2024/01/global-risks-report-2024/

Hellman, D. (2020). Measuring Algorithmic Fairness. *Virginia Law Review*, 106(4). https://virginialawreview.org/articles/measuring-algorithmic-fairness/

Henriksen, A., Enni, S., and Bechmann, A. (2021). Situated accountability: Ethical principles, certification standards, and explanation methods in applied ai. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 574–585. https://doi.org/10.1145/3461702.3462564

Herman, H. and Gornet, M. (2024). La normalisation de l'IA: un déluge de réinterprétations de l'AI Act. In *Journée Société et IA (SIA @ PFIA 2024)*. Association Française pour l'Intelligence Artificielle (AFIA). https://pfia2024.univ-lr.fr/Pr%C3%A9sentations/Soci%C3%A9t%C3%A9-et-IA---Pr%C3%A9sentation-3.1/

Hernalsteen, L. and Kohler, C. (2022). Drafting Harmonized Standards in support of the Artificial Intelligence Act (AIA) – CEN-CENELEC. https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN-CENELEC_Topics/Artificial%20Intelligence/jtc-21-harmonized-standards-webinar_for-website.pdf

Hesser, W. and Czaya, A. (2010). 12. Lecture: The Standardization Policy of the European Union by Wilfried Hesser & Axel Czaya. https://www.researchgate.net/publication/354047952_12_Lecture_The_Standardization_Policy_of_the_European_Union_by_Wilfried_Hesser_Axel_Czaya

Hidvegi, F. (2021). The EU should regulate AI on the basis of rights, not risks. *Access Now*. https://www.accessnow.org/eu-regulation-ai-risk-based-approach/

Hill, K. (2020a). Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. *The New York Times*. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Hill, K. (2020b). Wrongfully Accused by an Algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press. https://doi.org/10.1017/S0305004100013517

HLEG (2019a). Ethics guidelines for trustworthy AI. Technical report, Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission. Publications Office of the European Union, Directorate-General for Communications Networks, Content and Technology. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

HLEG (2019b). Policy and investment recommendations for trustworthy Artificial Intelligence. Technical report, Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/465913

HLEG (2020a). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Technical report, Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission. Publications Office of the European Union, Directorate-General for Communications Networks, Content and Technology. https://data.europa.eu/doi/10.2759/002360

HLEG (2020b). Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence. Technical report, Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission. Publications Office of the European Union, Directorate-General for Communications Networks, Content and Technology. https://data.europa.eu/doi/10.2759/733662

Ho-Dac, M. (2023). Considering Fundamental Rights in the European Standardisation of Artificial Intelligence: Nonsense or Strategic Alliance? In Jakobs, K., editor, *Joint Proceedings EURAS & SIIT 2023 – (Responsible) Standardisation for Smart Systems*. Verlag Mainz. https://hal.science/hal-04411136

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hofmann, H. C. H. (2016). A European Regulatory Union - The Role of Agencies and Standards. In Koutrakos, P. and Snell, J., editors, *Research Handbook on the EU's Internal Market*. University of Luxembourg Law Working Paper No. 2016-01. https://dx.doi.org/10.2139/ssrn.2745252

Hood, C., Rothstein, H., and Baldwin, a. R. (2001). *The Government of Risk: Understanding Risk Regulation Regimes*. Oxford University Press. https://global.oup.com/academic/product/the-government-of-risk-9780199270019

Horne, R. E. (2009). Limits to labels: The role of eco-labels in the assessment of product sustainability and routes to sustainable consumption. *International Journal of Consumer Studies*, 33(2):175–182. https://doi.org/10.1111/j.1470-6431.2009.00752.x

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hudson, J., Nemec, J., and Orviská, M. (2013). Standardization and the European Standards Organisations. *Central European Journal of Public Policy*, 7(02):36–58. Univerzita Karlova v Praze, Fakulta soci&#225;ln&#237;ch věd. https://www.ceeol.com/search/article-detail?id=264704

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Hunyadi, M. (2015). *La Tyrannie Des Modes De Vie. Sur Le Paradoxe Moral De Notre Temps*. Le Bord de l'Eau. https://www.editionsbdl.com/produit/la-tyrannie-des-modes-de-vie-sur-le-paradoxe-moral-de-notre-temps/

Hunyadi, M. (2018). Which Ethics for Bioethics? The Example of CRISPR-Cas9. Institut protestant de Paris, Faculté de Théologie. http://hdl.handle.net/2078.1/196134

Hutter, B. M. (2006). The role of non-state actors in regulation. pages 63–79. Nomos Verlagsgesellschaft. http://eprints.lse.ac.uk/id/eprint/9579

IAF (n.d.). Certification Bodies. International Accreditation Forum. https://www.iafcertsearch.org/search/certification-bodies

IEC (n.d.). National Committees. https://www.iec.ch/national-committees

IEEE (2021a). EEE Std 7000™-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design. IEEE Standards Association. https://standards.ieee.org/ieee/7000/6781/

IEEE (2021b). EEE Std 7001™-2021: IEEE Standard for Transparency of Autonomous Systems. IEEE Standards Association. https://standards.ieee.org/ieee/7001/6929/

IEEE (n.d.a). AIS Standards. IEEE Standards Association. https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/

IEEE (n.d.b). IEEE CertifAIEd™ Assessor Training. IEEE Standards Association. https://engagestandards.ieee.org/ieeecertifaied.html

IEEE (n.d.c). The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems. IEEE Standards Association. https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/

IGI Gloabl (n.d.). What is Algorithm. https://www.igi-global.com/dictionary/algorithms-aided-sustainable-urban-design/988

ISO (2008). ISO/IEC DIS 29500 receives necessary votes for approval as an International Standard. https://www.iso.org/news/2008/04/Ref1123.html

ISO (2010). ISO 26000:2010, Guidance on social responsibility. Edition 1. https://www.iso.org/standard/42546.html

ISO (2015a). ISO 14001:2015, Environmental management systems — Requirements with guidance for use. Edition 3. https://www.iso.org/standard/60857.html

ISO (2015b). ISO 9001:2015, Quality management systems — Requirements. Edition 5. https://www.iso.org/standard/62085.html

ISO (2021). ISO 37000:2021, Governance of organizations — Guidance. Edition 1. https://www.iso.org/standard/65036.html

ISO (2024). ISO House Style. https://www.iso.org/ISO-house-style.html

ISO (n.d.a). 2022 in review. https://www.iso.org/ar2022.html

ISO (n.d.b). ISO 14000 family: Environmental management. https://www.iso.org/standards/popular/iso-14000-family

ISO (n.d.c). ISO 9000 family: Quality management. https://www.iso.org/standards/popular/iso-9000-family

ISO (n.d.d). ISO/IEC JTC 1/SC 29 - Coding of audio, picture, multimedia and hypermedia information. https://www.iso.org/committee/45316.html

ISO (n.d.e). ISO/IEC JTC 1/SC 37 - Biometrics. https://www.iso.org/committee/313770.html

ISO (n.d.f). ISO/IEC JTC 1/SC 42 - Artificial intelligence. https://www.iso.org/committee/6794475.html

ISO (n.d.g). Members. https://www.iso.org/about/members

ISO (n.d.h). Standards. https://www.iso.org/standards.html

ISO (n.d.i). Who develops standards. https://www.iso.org/who-develops-standards.html

ISO (n.d.j). Young professionals in standards. https://www.iso.org/strategy2030/young-professionals-in-standards

ISO and CEN (2016a). Vienna Agreement 7th edition – Guidelines for the implementation of the Agreement on Technical Cooperation between ISO and CEN. https://boss.cen.eu/media/CEN/ref/va_guidelines_implementation.pdf

ISO and CEN (2016b). The Vienna Agreement – FAQs. https://boss.cen.eu/media/CEN/ref/va_faq.pdf

ISO/IEC (1991). ISO/IEC Guide 2:1991, General terms and their definitions concerning standardization and related activities. Edition 6. https://www.iso.org/standard/19713.html

ISO/IEC (2005). ISO/IEC 27001:2005, Information technology — Security techniques — Information security management systems — Requirements. Edition 1. https://www.iso.org/contents/data/standard/04/21/42103.html

ISO/IEC (2015). ISO/IEC 17021-1:2015, Conformity assessment — Requirements for bodies providing audit and certification of management systems. Part 1: Requirements. Edition 1. https://www.iso.org/standard/61651.html

ISO/IEC (2017). ISO/IEC 17011:2017, Conformity assessment — Requirements for accreditation bodies accrediting conformity assessment bodies. Edition 2. https://www.iso.org/standard/67198.html

ISO/IEC (2019). ISO/IEC 27701:2019, Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines. Edition 1. https://www.iso.org/standard/71670.html

ISO/IEC (2020). ISO/IEC TR 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. Edition 1. https://www.iso.org/standard/77608.html

ISO/IEC (2021). ISO/IEC TR 24027:2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making. Edition 1. https://www.iso.org/standard/77607.html

ISO/IEC (2022a). ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems — Requirements. Edition 3. https://www.iso.org/standard/27001

ISO/IEC (2022b). ISO/IEC 27002:2022, Information security, cybersecurity and privacy protection — Information security controls. Edition 3. https://www.iso.org/standard/75652.html

ISO/IEC (2022c). ISO/IEC 38507:2022, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. Edition 1. https://www.iso.org/standard/56641.html

ISO/IEC (2022d). ISO/IEC TR 24368:2022, Information technology — Artificial intelligence — Overview of ethical and societal concerns. Edition 1. https://www.iso.org/standard/78507.html

ISO/IEC (2023a). ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management. Edition 1. https://www.iso.org/standard/77304.html

ISO/IEC (2023b). ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system. Edition 1. https://www.iso.org/standard/81230.html

ISO/IEC (2024a). ISO/IEC 19795-10:2024, Information technology — Biometric performance testing and reporting. Part 10: Quantifying biometric system performance variation across demographic groups. Edition 1. https://www.iso.org/standard/81223.html

ISO/IEC (2024b). ISO/IEC 5259-1:2024, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples. Edition 1. https://www.iso.org/standard/81088.html

ISO/IEC (n.d.a). ISO/IEC 27000 family: Information security management. https://www.iso.org/standard/iso-iec-27000-family

ISO/IEC (n.d.b). ISO/IEC AWI TSI 22443, Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations. https://www.iso.org/standard/87119.html

ISO/IEC (n.d.c). ISO/IEC CD TS 6254, Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems. https://www.iso.org/standard/82148.html

ISO/IEC (n.d.d). ISO/IEC DIS 42006, Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems. https://www.iso.org/standard/44546.html

ISO/IEC (n.d.e). ISO/IEC TS 12791, Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks. https://www.iso.org/standard/84110.html

ITEH Standards (n.d.). Mandate M/530 - Privacy Management. https://standards.iteh.ai/catalog/mandate/cen/90507928-bb87-4b60-8c56-fdd18d6dd2db/m-530

ITU (2022). ITU-T Y.3000-series – Artificial intelligence standardization roadmap. ITU-T Y Suppl. 72. SERIES Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities. https://www.itu.int/ITU-T/recommendations/rec.aspx?id=15249&lang=en

ITU (n.d.). Membership. https://www.itu.int/hub/membership/our-members/

JBCE (2021). The Roadmap of European Standardisation Strategy. Japan Business Council in Europe. https://www.old.jbce.org/standards-and-conformity/jbce-input-to-the-roadmap-of-european-standardisation-strategy/

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399. Nature Publishing Group. https://doi.org/10.1038/s42256-019-0088-2

John-Mathews, J.-M. (2021). *AI ethics in practice, challenges and limitations*. PhD thesis. https://www.researchgate.net/publication/357909990_AI_ethics_in_practice_challenges_and_limitations

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169. https://doi.org/10.1038/d41586-020-02003-2

Kamara, I. (2017). Co-regulation in EU personal data protection: the case of technical standards and the privacy by design standardisation 'mandate'. *European Journal of Law and Technology*, 8(1). https://ejlt.org/index.php/ejlt/article/view/545

Kamara, I. (2021). *Data protection standardisation: The role and limits of technical standards in the EU data protection law*. PhD thesis, Tilburg University.

Kaminski, M. E. (2022). Regulating the Risks of AI. *Boston University Law Review, U of Colorado Law Legal Studies*, 103(1347). https://dx.doi.org/10.2139/ssrn.4195066

Kanevskaia, O. (2024). The Commission strikes back: regulatory authority and legitimacy in European technology standardization. A commentary on Cantero Gamito and Kamara. *Innovation: The European Journal of Social Science Research*. https://doi.org/10.1080/13511610.2024.2332235

Kaplinsky, R. (2010). The Role of Standards in Global Value Chains. *World Bank Policy Research Working Paper No. 5396*. https://ssrn.com/abstract=1653682

Kayser-Bril, N. (2020). Google apologizes after its Vision AI produced racist results. AlgorithmWatch. https://algorithmwatch.org/en/google-vision-racism/

Keller, J. and Levallois-Barth, C. (2021). La fragile définition de l'identité européenne par ses valeurs numériques. *Revue générale du droit*, (57745). Chronique de droit de l'Union. www.revuegeneraledudroit.eu/?p=57745

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., and Akbar, M. A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering 2022*, EASE '22, pages 383–392. Association for Computing Machinery. https://doi.org/10.1145/3530019.3531329

Kim, P. T. (2022). Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action. *California Law Review*, 110. https://www.californialawreview.org/print/race-aware-algorithms-fairness-nondiscrimination-and-affirmative-action

Kirk, J. (2008). ISO publishes Office Open XML specification. *InfoWorld*. https://www.infoworld.com/article/2178151/iso-publishes-office-open-xml-specification.html

Knight, W. (2019a). About 40% of Europe's "AI companies" don't use any AI at all. *MIT Technology Review*. https://www.technologyreview.com/2019/03/05/65990/about-40-of-europes-ai-companies-dont-actually-use-any-ai-at-all/

Knight, W. (2019b). The Apple Card Didn't "See" Gender—and That's the Problem. *Wired*. https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Koene, A., Dowthwaite, L., and Seth, S. (2018). IEEE P7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness*, pages 38–41. Association for Computing Machinery. https://doi.org/10.1145/3194770.3194773

Kohli, P. and Chadha, A. (2019). Enabling Pedestrian Safety Using Computer Vision Techniques: A Case Study of the 2018 Uber Inc. Self-driving Car Crash. In Arai, K. and Bhatia, R., editors, *Advances in Information and Communication. FICC 2019. Lecture Notes in Networks and Systems*, volume 69, pages 261–279. Springer International Publishing. https://doi.org/10.1007/978-3-030-12388-8_19

Kozyrkov, C. (2024). Why AI and decision-making are two sides of the same coin. *Medium*. https://towardsdatascience.com/in-ai-the-objective-is-subjective-4614795d179b

Krishnan, A., Abdilla, A., Moon, A. J., Souza, C. A., Adamson, C., Lach, E. M., Ghazal, F., Fjeld, J., Taylor, J., Havens, J. C., Jayaram, M., Morrow, M., Rizk, N., Quijano, P. R., Çetin, R. B., Chatila, R., Dotan, R., Mhlambi, S., Jordan, S., and Rosenstock, S. (n.d.). Decolonial AI Manyfesto. https://manyfesto.ai/

Kubrick, S. (1968). 2001: A Space Odyssey. [Film].

Lachaud, E. (2018). The General Data Protection Regulation and the rise of certification as a regulatory instrument. *Computer Law & Security Review*, 34(2):244–256. https://doi.org/10.1016/j.clsr.2017.09.002

Lachaud, E. (2020). What GDPR tells about certification. *Computer Law & Security Review*, 38. https://doi.org/10.1016/j.clsr.2020.105457

Lambrecht, A. and Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7):2966–2981. INFORMS. https://pubsonline.informs.org/doi/10.1287/mnsc.2018.3093

Langlois, L. and Régis, C. (2021). Analyzing the Contribution of Ethical Charters to Building the Future of Artificial Intelligence Governance. In Braunschweig, B. and Ghallab, M., editors, *Reflections on Artificial Intelligence for Humanity*, Lecture Notes in Computer Science, pages 150–170. Springer International Publishing. https://doi.org/10.1007/978-3-030-69128-8_10

Larson, D. B. and Jordan, S. R. (2019). Playing it safe: toy safety and conformity assessment in Europe and the United States. *International Review of Administrative Sciences*, 85(4):763–779. SAGE Publications. https://doi.org/10.1177/0020852317747370

Laux, J., Wachter, S., and Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32. John Wiley & Sons. https://doi.org/10.1111/rego.12512

Laux, J., Wachter, S., and Mittelstadt, B. (2024). Three pathways for standardisation and ethical disclosure by default under the European union artificial intelligence act. *Computer Law & Security Review*, 53. Elsevier Advanced Technology. https://doi.org/10.1016/j.clsr.2024.105957

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343:1203–1205. https://tinyurl.com/y3r58os2

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1:541–551. https://yann.lecun.com/exdb/publis/pdf/lecun-89e.pdf

Leffrang, D. and Mueller, O. (2023). AI Washing: The Framing Effect of Labels on Algorithmic Advice Utilization. *International Conference on Information Systems (ICIS) 2023 Proceedings*, (10). https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/10

Liboiron, M. (2021). *Pollution is colonialism*. Duke University Press. https://www.dukeupress.edu/pollution-is-colonialism

LNE (2021). Certification Standard of Processes for AI Design, development, evaluation and maintenance in operational conditions. Laboratoire National de Métrologie et d'Essais. https://www.lne.fr/en/service/certification/certification-processes-ai

Lohr, S. (2024). A.I. Can Write Poetry, but It Struggles With Math. *The New York Times*. https://www.nytimes.com/2024/07/23/technology/ai-chatbots-chatgpt-math.html

Lopes, I. M., Guarda, T., and Oliveira, P. (2019). How ISO 27001 Can Help Achieve GDPR Compliance. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. https://doi.org/10.23919/CISTI.2019.8760937

Loya, T. A. and Boli, J. (1999). Standardization in the World Polity: Technical Rationality Over Power. In Boli, J. and Thomas, G. M., editors, *Constructing world culture : international nongovernmental organizations since 1875*. Stanford University Press. https://www.sup.org/books/title/?id=2040

Ludec, C. L., Cornet, M., and Casilli, A. A. (2023). The problem with annotation. human labour and outsourcing between france and madagascar. *Big Data & Society*, 10(2). https://doi.org/10.1177/20539517231188723

Lundqvist, B. (2014). *Standardization under EU Competition Rules and US Antitrust Laws*. Law 2014. Edward Elgar Publishing. https://doi.org/10.4337/9781781954867

Légifrance (2009). Décret n° 2009-697 du 16 juin 2009 relatif à la normalisation. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000020749979

Madiega, T. (2023). Artificial intelligence liability directive. *European Parliament Briefing*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf

Marchant, G. E. (2011). Addressing the Pacing Problem. In Marchant, G. E., Allenby, B. R., and Herkert, J. R., editors, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, volume 7, pages 199–205. Springer Netherlands, springer edition. https://doi.org/10.1007/978-94-007-1356-7_13

Marcinkowski, A. and Wilgaux, J. (2004). Automates et créatures artificielles d'Héphaïstos : entre science et fiction. *Techniques & Culture. Revue semestrielle d'anthropologie des techniques*, (43-44). https://doi.org/10.4000/tc.1164

Martinez-Julia, P., Homma, S., and Lopez, D. (2021). Artificial Intelligence Framework for Network Management. Internet Engineering Task Force. https://datatracker.ietf.org/doc/draft-pedro-nmrg-ai-framework-00

Maruf, R. (2022). Google AI is real, says fired engineer. *CNN*. https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html

Mattli, W. and Buthe, T. (2003). Setting International Standards: Technological Rationality or Primacy of Power? *World Politics*, 56(1):1–42. Johns Hopkins University Press. https://doi.org/10.1353/wp.2004.0006

Matutes, C. and Regibeau, P. (1996). A selective review of the economics of standardization. Entry deterrence, technological progress and international competition. *European Journal of Political Economy*, 12(2):183–209. https://doi.org/10.1016/0176-2680(95)00013-5

Maxwell, W. (2022). Les modes de régulation des activités numériques : exploration des tensions entre l'approche par les risques (risk-based) et l'approche fondée sur la protection des droits. Université Paris 1 Panthéon- Sorbonne. https://hal.science/tel-04026744

Mazzini, G. and Scalzo, S. (2023). The Proposal for the Artificial Intelligence Act: Considerations around Some Key Concepts. In Camardi, C., editor, *La via europea per l'Intelligenza artificiale*. Cedam. https://dx.doi.org/10.2139/ssrn.4098809

McFadden, M., Jones, K., Taylor, E., and Osborn, G. (2021). Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation. *Oxford Commission on AI & Good Governance*. Oxford Information Labs. https://www.oii.ox.ac.uk/news-events/reports/harmonising-artificial-intelligence/

McNamara, A., Smith, J., and Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, pages 729–733, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3236024.3264833

Merriam-Webster Dictionary (n.d.a). Anthropomorphic. https://www.merriam-webster.com/dictionary/anthropomorphic

Merriam-Webster Dictionary (n.d.b). Artificial intelligence. https://www.merriam-webster.com/dictionary/artificial%20intelligence

Merriam-Webster Dictionary (n.d.c). Autonomy. https://www.merriam-webster.com/dictionary/autonomy

Merriam-Webster Dictionary (n.d.d). Bias. https://www.merriam-webster.com/dictionary/bias

Merriam-Webster Dictionary (n.d.e). Consensus. https://www.merriam-webster.com/dictionary/consensus

Merriam-Webster Dictionary (n.d.f). Data. https://www.merriam-webster.com/dictionary/data

Merriam-Webster Dictionary (n.d.g). Ethics. https://www.merriam-webster.com/dictionary/ethics

Merriam-Webster Dictionary (n.d.h). Ethnicity. https://www.merriam-webster.com/dictionary/ethnicity

Merriam-Webster Dictionary (n.d.i). Regulate. https://www.merriam-webster.com/dictionary/regulate

Merriam-Webster Dictionary (n.d.j). Standard. https://www.merriam-webster.com/dictionary/standard

Metz, C. (2023). Chatbots May 'Hallucinate" More Often Than Many Realize. *The New York Times*. https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

Metzinger, T. (2019). Eu guidelines: Ethics washing made in Europe. *Tagesspiegel*. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Microsoft (2021). Microsoft's Response to the European Commission's Consultation on the Artificial Intelligence Act. European Commission - Have your say. Feedback from: Microsoft Corporation. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665556_en

Microsoft (2022). Microsoft Responsible AI Standard, v2. General Requirements. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Miotti, H. (2009). The Economic impact of Standardization - Technological chance, standards growth in France. Technical report, AFNOR. https://normalisation.afnor.org/wp-content/uploads/2016/06/Etude-ImpactEcoNorm-GB2009.pdf

Mitnick, B. M. (2011). Capturing "Capture": Definition and Mechanisms. In Levi-Faur, D., editor, *Handbook on the Politics of Regulation*. Edward Elgar Publishing. https://doi.org/10.4337/9780857936110.00013

Mittelstadt, B. (2019). Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*. https://dx.doi.org/10.2139/ssrn.3391293

Moravec, H. (1990). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press. https://www.hup.harvard.edu/books/9780674576186

Morgan, B. and Yeung, K. (2007). *An Introduction to Law and Regulation: Text and Materials*. Law in Context. Cambridge University Press. https://doi.org/10.1017/CBO9780511801112

Morikawa, M. and Morrison, J. (2004). Who Develops ISO Standards? A Survey of Participation in ISO's International Standards Development Processes. Technical report, Pacific Institute for Studies in Development, Environment, and Security. https://pacinst.org/publication/who-develops-iso-standards-report-surveys-participation-in-standards-creation/

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., and Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of ai ethics. *Minds and Machines*, 31(2):239–256. Springer. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 26(4):2141–2168. Springer. https://doi.org/10.1007/s11948-019-00165-5

Mueck, M., Cadzow, S., Communications, C., and Wood, S. (2022). ETSI White Paper No. #52. ETSI Activities in the field of Artificial Intelligence Preparing the implementation of the European AI Act. 1st Edition. Technical report, European Telecommunications Standards Institute (ETSI). https://www.etsi.org/images/files/ETSIWhitePapers/ETSI-WP52-ETSI-activities-in-the-field-of-AI.pdf

Mulligan, D. K., Kroll, J. A., Kohli, N., and Wong, R. Y. (2019). This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW, article 119). https://doi.org/10.1145/3359221

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 3. Springer. https://doi.org/10.1007/s43681-022-00209-w

Mönnig, R. and Manouchehri, S. (2024). Rethinking Europe's Digital Future: Challenges and Risks from GAFAM. In van de Wetering, R., Helms, R., Roelens, B., Bagheri, S., Dwivedi, Y. K., Pappas, I. O., and Mäntymäki, M., editors, *Disruptive Innovation in a Digitally Connected Healthy World*, pages 241–252. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72234-9_20

Müller, J., Mitesser, O., Schaefer, H. M., Seibold, S., Busse, A., Kriegel, P., Rabl, D., Gelis, R., Arteaga, A., Freile, J., Leite, G. A., de Melo, T. N., LeBien, J., Campos-Cerqueira, M., Blüthgen, N., Tremlett, C. J., Böttger, D., Feldhaar, H., Grella, N., Falconí-López, A., Donoso, D. A., Moriniere, J., and Buřivalová, Z. (2023). Soundscapes and deep learning enable tracking biodiversity recovery in tropical forests. *Nature Communications*, 14(1):6191. https://doi.org/10.1038/s41467-023-41693-w

Narayanan, A. and Kapoor, S. (2024). *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691249131/ai-snake-oil

Nasir, O., Muntaha, S., Javed, R. T., and Qadir, J. (2021). Work in Progress: Pedagogy of Engineering Ethics: A Bibliometric and Curricular Analysis. In *2021 IEEE Global Engineering Education Conference (EDUCON)*, pages 1553–1557. https://ieeexplore.ieee.org/document/9453923

Nativi, S. and De Nigris, S. (2021). AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework. Technical report, European Commission Joint Research Centre. Publications Office of the European Union. https://dx.doi.org/10.2760/376602

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. In Cath, C., Wachter, S., Mittelstadt, B., and Floridi, L., editors, *Governing artificial intelligence: ethical, legal, and technical opportunities and challenges*, volume 376. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://doi.org/10.1098/rsta.2018.0089

NEN (2021). Statement on the European Commission's Roadmap for a Standardization Strategy. Executive Board of the Royal Netherlands Standardization Institute (NEN). https://intranet.nen.nl/wp-content/uploads/2021/08/Statement-NEN-on-the-2021-EC-Roadmap-Standardization-Strategy-Final.pdf

Nickel, P. J., Franssen, M., and Kroes, P. (2010). Can We Make Sense of the Notion of Trustworthy Technology? *Knowledge, Technology & Policy*, 23(3):429–444. https://doi.org/10.1007/s12130-010-9124-6

NIST (2023a). AI Risk Management Framework: AI RMF (1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

NIST (2023b). NIST AI Program. Artificial Intelligence: The Vitals. https://www.nist.gov/system/files/documents/2023/11/02/AI%20Fact%20Sheet%200615%20FINAL.pdf

NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Technical Report NIST AI NIST AI 600-1, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

NIST (n.d.a). Face Recognition Technology Evaluation (FRTE) 1:1 Verification. https://pages.nist.gov/frvt/html/frvt11.html

NIST (n.d.b). Face Recognition Vendor Test (FRVT). https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

Northpointe (2019). Practitioner's Guide to COMPAS Core. https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf

NSTC (2016). Preparing for the future of AI - Executive Office of the President National Science and Technology Council Committee on Technology. Technical report, National Science and Technology Council (US). https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

OECD (2019). Recommendation of the Council on Artificial Intelligence. C/MIN(2019)3/FINAL. https://one.oecd.org/document/C/MIN(2019)3/FINAL/en/pdf

OECD (2024a). AI, data governance and privacy: Synergies and areas of international co-operation. Technical report, OECD Artificial Intelligence papers. https://www.oecd.org/en/publications/ai-data-governance-and-privacy_2476b1a4-en.html

OECD (2024b). Recommendation of the Council on Artificial Intelligence. C/MIN(2024)16/FINAL. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD (n.d.). The OECD Artificial Intelligence Policy Observatory: Policies, data and analysis for trustworthy artificial intelligence. https://oecd.ai/en/

Oever, N. t. and Milan, S. (2022). The Making of International Communication Standards: Towards a Theory of Power in Standardization. *Journal of Standardisation*, 1. https://doi.org/10.18757/jos.2022.6205

OHCHR (2023). Human rights and technical standard-setting processes for new and emerging digital technologies : Report of the Office of the United Nations High Commissioner for Human Rights. Office of the United Nations High Commissioner for Human Rights. https://digitallibrary.un.org/record/4031373

OJEU (2025). Action brought on 6 December 2024 – International Electrotechnical Commission and ISO v Commission (Case T-631/24). Official Journal of the European Union (OJ C, C/2025/919). http://data.europa.eu/eli/C/2025/919/oj

OMB (1998). Circular n°A-119 Revised. Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-119-1.pdf

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA. https://dl.acm.org/doi/10.5555/3002861

Orwat, C., Bareis, J., Folberth, A., Jahnel, J., and Wadephul, C. (2024). Normative Challenges of Risk Regulation of Artificial Intelligence. *Nanoethics*, 18(11). https://doi.org/10.1007/s11569-024-00454-9

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://arxiv.org/abs/2203.02155

Papadopoulos, T. and Charalabidis, Y. (2020). What do governments plan in the field of artificial intelligence?: Analysing national AI strategies using NLP. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, pages 100–111. Association for Computing Machinery. https://doi.org/10.1145/3428502.3428514

Papakonstantinou, V. and De Hert, P. (2024). *The Regulation of Digital Technologies in the EU: Act-ification, GDPR Mimesis and EU Law Brutality at Play*. Routledge. https://www.routledge.com/The-Regulation-of-Digital-Technologies-in-the-EU-Act-ification-GDPR-Mimesis-and-EU-Law-Brutality-at-Play/Papakonstantinou-DeHert/p/book/9781032614441

Papyshev, G. and Yarime, M. (2023). The state's role in governing artificial intelligence: development, control, and promotion through national strategies. *Policy Design and Practice*, pages 1–24. https://doi.org/10.1080/25741292.2022.2162252

Partnership on AI (2016). Our Tenets. https://partnershiponai.org/about/#tenets

Paul, R. and Loh, E. (2023). Europe's Blue Guide, New Approach and New Legislative Framework. *Emergo by UL*. https://www.emergobyul.com/news/europes-blue-guide-new-approach-and-new-legislative-framework

Pelkmans, J. (1987). The New Approach to Technical Harmonization and Standardization. *Journal of Common Market Studies*, 25(3):249–269. https://doi.org/10.1111/j.1468-5965.1987.tb00294.x

Perarnaud, C. (2023). With the AI Act, we need to mind the standards gap. *Centre for European Policy Studies*. https://www.ceps.eu/with-the-ai-act-we-need-to-mind-the-standards-gap/

Perez, C. C. (2020). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage Books, London. https://carolinecriadoperez.com/book/invisible-women/

Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Peylo, C., Slama, D., Hallensleben, S., Hauschke, A., and Hildebrandt, S. (2022). VCIO based description of systems for AI trustworthiness characterisation. Technical Report VDE SPEC 90012 V1.0 (en), Verband der Elektrotechnik (VDE). https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf

Pfungst, O. (1911). *Clever Hans (the horse of Mr. Von Osten) a contribution to experimental animal and human psychology*. Henry Holt and Company. E-book version by Project Gutenberg. Edition 2010. https://www.gutenberg.org/files/33936/33936-h/33936-h.htm

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661. JSTOR. https://www.jstor.org/stable/1806107

Pinch, T. J. and Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, 14(3):399–441. SAGE Publications. https://doi.org/10.1177/030631284014003004

Pouget, H. (2023). The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist. *Lawfare*. https://www.lawfaremedia.org/article/eus-ai-act-barreling-toward-ai-standards-do-not-exist

Pouget, H. (2024). Standard Setting. Future of Life Institute. https://artificialintelligenceact.eu/standard-setting/

Pouget, H. and Laux, J. (2023). A Letter to the EU's Future AI Office. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/research/2023/10/a-letter-to-the-eus-future-ai-office?lang=en

Probasco, E., Toney, A., and Curlee, K. (2023). The Inigo Montoya Problem for Trustworthy AI. Technical report, Center for Security and Emerging Technology. https://doi.org/10.51593/20230014a

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v

Reardon, T., Codron, J.-M., Busch, L., Bingen, R. J., and Harris, C. (1999). Global change in agrifood grades and standards: agribusiness strategic responses in developing countries. *International Food and Agribusiness Management Review*, 02(3-4). http://dx.doi.org/10.22004/ag.econ.34227

Rees, C. and Müller, B. (2022). All that glitters is not gold: trustworthy and ethical AI principles. *AI and Ethics*, 3. Springer. https://doi.org/10.1007/s43681-022-00232-x

Reinert, M. (1990). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les cahiers de l'analyse des données*, 15(1):21–36. http://www.numdam.org/item/CAD_1990__15_1_21_0/

Reséguier, A. and Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2). SAGE Publications. https://doi.org/10.1177/2053951720942541

Rezk, E., Eltorki, M., and El-Dakhakhni, W. (2022). Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach. *JMIR Dermatology*, 5(3). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10334920/

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *arXiv preprint arXiv:1602.04938v3*. https://arxiv.org/abs/1602.04938v3

Rieder, G., Simon, J., and Wong, P.-H. (2021). Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. In Pelillo, M. and Scantamburlo, T., editors, *Machines We Trust: Perspectives on Dependable AI*. MIT Press Direct. https://doi.org/10.7551/mitpress/12186.003.0007

Rivera Tello, G. A., Takahashi, K., and Karamperidou, C. (2023). Explained predictions of strong eastern Pacific El Niño events using deep learning. *Scientific Reports*, 13(1):21150. Nature. https://doi.org/10.1038/s41598-023-45739-3

Robotics Openletter (2017). Open letter to the European Commission – Artificial Intelligence and Robotics. https://robotics-openletter.eu/

Roche, C., Wall, P. J., and Lewis, D. (2022). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, 3. Springer. https://doi.org/10.1007/s43681-022-00218-9

Rodrigues, R., Barnard-Wills, D., Wright, D., De Hert, P., and Papakonstantinou, V. (2013). *EU privacy seals project: inventory and analysis of privacy certification schemes*. Publications Office of the European Union. JRC85092. https://dx.doi.org/10.2788/29861

Rothstein, H., Irving, P., Walden, T., and Yearsley, R. (2006). The risks of risk-based regulation: Insights from the environmental policy domain. *Environment International*, 32(8):1056–1065. https://doi.org/10.1016/j.envint.2006.06.008

Rott, P. (2019). Certification of Medical Devices: Lessons from the PIP Scandal. In Rott, P., editor, *Certification – Trust, Accountability, Liability*, volume 16, pages 189–211. Springer International Publishing, Studies in European Economic Law and Regulation. https://doi.org/10.1007/978-3-030-02499-4_9

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*. https://www.nature.com/articles/s42256-019-0048-x

Ruschemeier, H. (2023). AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum*, 23(3):361–376. https://doi.org/10.1007/s12027-022-00725-6

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, S. (2021). Roadmap for the standardisation strategy. ANEC response. Technical report, The European Association for the Co-ordination of Consumer Representation in Standardisation (ANEC). https://anec.eu/images/2021/ANEC-WP1-2021-046.pdf

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5):2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Ryan, M. and Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, 19(1):61–86. Emerald Publishing Limited. https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html

Saheb, T., Dehghani, M., and Saheb, T. (2022). Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. *Sustainable Computing: Informatics and Systems*, 35. https://doi.org/10.1016/j.suscom.2022.100699

Sahouri, A. M. (2024). Man wrongfully arrested by Detroit police with facial recognition tech settles lawsuit. *Detroit Free Press*. https://www.freep.com/story/news/local/michigan/detroit/2024/06/28/man-wrongfully-arrested-with-facial-recognition-tech-settles-lawsuit/74243839007/

Saltelli, A., Dankel, D. J., Di Fiore, M., Holland, N., and Pigeon, M. (2022). Science, the endless frontier of regulatory capture. *Futures*, 135. Pergamon. https://doi.org/10.1016/j.futures.2021.102860

SAP (2024). Global Artificial Intelligence (AI) Ethics Policy. https://www.sap.com/documents/2022/01/a8431b91-117e-0010-bca6-c68f7e60039b.html

Schapel, H. (2013). The New Approach to the New Approach: The Juridification of Harmonized Standards in EU Law. *Maastricht Journal of European and Comparative Law*, 20(4):521–533. SAGE Publications. https://doi.org/10.1177/1023263X1302000404

Schepel, H. (2005). *The constitution of private governance: product standards in the regulation of integrating markets*. Number 4 in International studies in the theory of private law. Hart Publishing. https://doi.org/10.5040/9781472563255

Scheuerman, M. K., Paul, J. M., and Brubaker, J. R. (2019). How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33. https://dl.acm.org/doi/10.1145/3359246

Schischke, K., Berwald, A., Dimitrova, G., Rückschloss, J., Nissen, N. F., and Schneider-Ramelow, M. (2022). Durability, reparability and recyclability: Applying material efficiency standards EN 4555x to mobile phones and tablet computers. *Procedia CIRP*, 105:619–624. https://doi.org/10.1016/j.procir.2022.02.103

Schmidt, S. K. and Werle, R. (1998). *Coordinating Technology – Studies in the International Standardization of Telecommunications*. MIT Press. https://mitpress.mit.edu/9780262193931/coordinating-technology/

Schmidt, V. A. (2013). Democracy and Legitimacy in the European Union Revisited: Input, Output and 'Throughput'. *Political Studies Association*, 61(1):2–22. SAGE Publications. https://doi.org/10.1111/j.1467-9248.2012.00962.x

Scholten, M. and van Rijsbergen, M. (2014). The ESMA-short selling case: Erecting a new delegation doctrine in the EU upon the Meroni-Romano remnants. *Legal issues of European integration*, 41(4):389–405. Kluwer Law International. https://doi.org/10.54648/leie2014022

Schröder, S., Schulz, A., Kenneweg, P., Feldhans, R., Hinder, F., and Hammer, B. (2021). Evaluating metrics for bias in word embeddings. *arXiv preprint arXiv:2111.07864*. https://arxiv.org/abs/2111.07864

Schuett, J. (2023a). Defining the scope of AI regulations. *Law, Innovation and Technology*, 15(1):60–82. https://doi.org/10.1080/17579961.2023.2184135

Schuett, J. (2023b). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, pages 1–19. Cambridge University Press. https://doi.org/10.1017/err.2023.1

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424. https://doi.org/10.1017/S0140525X00005756

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68. https://doi.org/10.1145/3287560.3287598

Senado Federal (2023). Projeto de Lei 2338/2023. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

Senden, L. (2004). *Soft Law in European Community Law*. Modern Studies in European Law. Hart Publishing, bloomsbury publishing edition. https://www.bloomsbury.com/uk/soft-law-in-european-community-law-9781847311108/

Senden, L. (2017). The Constitutional Fit of European Standardization Put to the Test. *Legal Issues of Economic Integration*, 44(4). https://doi.org/10.54648/leie2017018

Senden, L. (2020). Towards a More Holistic Legitimacy Approach to Technical Standardisation in the EU. In Eliantonio, M. and Cauffman, C., editors, *The Legitimacy of Standardisation as a Regulatory Technique*, pages 20–47. Edward Elgar Publishing. https://doi.org/10.4337/9781789902952.00007

Senden, L. A. J. (2005). Soft Law, Self-Regulation and Co-Regulation in European Law: Where Do They Meet? *Electronic Journal of Comparative Law*, 9(1). https://ssrn.com/abstract=943063

Shapiro, C. and Varian, H. R. (1999). The Art of Standards Wars. *California Management Review*, 41(2):8–32. SAGE Publications. https://doi.org/10.2307/41165984

Sharma, C. (2024). AI's Hippocratic Oath. *Washington University Law Review*. https://ssrn.com/abstract=4759742

Shelley, M. W. (1993). *Frankenstein; or, the Modern Prometheus*. E-book version by Project Gutenberg. Edition 1993. https://www.gutenberg.org/files/84/84-h/84-h.htm

Shepardson, D. (2025). Trump revokes Biden executive order on addressing AI risks. *Reuters*. https://www.reuters.com/technology/artificial-intelligence/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/

Siegmann, C. and Anderljung, M. (2022). The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market. *arXiv preprint arXiv:2208.12645*. https://arxiv.org/abs/2208.12645

Smuha, N. A. (2019). The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. *Computer Law Review International*, 20(4):97–106. Verlag Dr. Otto Schmidt. https://doi.org/10.9785/cri-2019-200402

Smuha, N. A. (2021a). Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philosophy & Technology*, 34(1):91–104. https://doi.org/10.1007/s13347-020-00403-w

Smuha, N. A. (2021b). From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence. https://doi.org/10.1080/17579961.2021.1898300

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., and Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. https://dx.doi.org/10.2139/ssrn.3899991

Smuha, N. A. and Yeung, K. (2024). The European Union's AI Act: beyond motherhood and apple pie? https://dx.doi.org/10.2139/ssrn.4874852

Soler Garrido, J., De Nigris, S., Bassani, E., Sanchez, I., Evas, T., André, A.-A., and Boulangé, T. (2024). Harmonised Standards for the European AI Act. *JRC Publications Repository*. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430

Soler Garrido, J., Fano Yela, D., Panigutti, C., Junklewitz, H., Hamon, R., Evas, T., André, A.-A., and Scalzo, S. (2023a). Analysis of the preliminary AI standardisation work plan in support of the AI Act. Technical report, European Commission Joint Research Centre. Publications Office of the European Union. https://dx.doi.org/10.2760/5847

Soler Garrido, J., Tolan, S., Hupont Torres, I., Fernández Llorca, D., Charisi, V., Gómez Gutiérrez, E., Junklewitz, H., Hamon, R., Fano Yela, D., and Panigutti, C. (2023b). AI Watch: artificial intelligence standardisation landscape update. Technical report, European Commission Joint Research Centre. Publications Office of the European Union. https://dx.doi.org/10.2760/131984

Solow-Niederman, A. (2024). Can AI Standards Have Politics? *UCLA Law Review*. `https://www.uclalawreview.org/can-ai-standards-have-politics/`

Soroiu, A. (2024). The Fall of The Great Paywall for EU Harmonised Standards. *Verfassungsblog*. `https://doi.org/10.59704/5a60ea5d42c2b059`

Spivak, S. M. and Brenner, F. C. (1993). *Standardization Essentials: Principles and Practice*. CRC Press. `https://doi.org/10.1201/9781482277388`

StandICT.eu (2023a). StandICT.eu 2026 - 1st Open Call - CLOSED. `https://standict.eu/standicteu-2026-1st-open-call`

StandICT.eu (2023b). StandICT.eu 2026 - 2nd Open Call. `https://standict.eu/standicteu-2026-2nd-open-call`

StandICT.eu (2024a). StandICT.eu 2026 - 3rd Open Call. `https://standict.eu/standicteu-2026-3rd-open-call`

StandICT.eu (2024b). StandICT.eu 2026 - 4th Open Call. `https://standict.eu/standicteu-2026-4th-open-call`

StandICT.eu (n.d.). StandICT.eu 2026. `https://standict.eu/standicteu-2026`

Stango, V. (2004). The Economics of Standards Wars. *Review of Network Economics*, 3(1). De Gruyter. `https://doi.org/10.2202/1446-9022.1040`

Staniszewski, P. (2007). The interplay between IP rights and competition law in the context of standardization. *Journal of Intellectual Property Law & Practice*, 2(10):666–681. `https://doi.org/10.1093/jiplp/jpm143`

Stanojevic, A. (2024). ETSI As a Case Study of Organizational Resilience in Standard Setting: an Empirical Inquiry Into Strategies That Ensure Thriving Despite Organizational Challenges. *Innovation: The European Journal of Social Science Research*. `https://doi.org/10.1080/13511610.2024.2369180`

Steffek, J. and Wegmann, P. (2021). The Standardization of "Good Governance" in the Age of Reflexive Modernity. *Global Studies Quarterly*, 1(4). `https://doi.org/10.1093/isagsq/ksab029`

Streeck, W. (1995). Neo-Voluntarism: A New European Social Policy Regime? *European Law Journal*, 1(1):31–59. `https://doi.org/10.1111/j.1468-0386.1995.tb00003.x`

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243*. `https://arxiv.org/abs/1906.02243`

Suresh, H. and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*. `https://hdl.handle.net/1721.1/143588`

Swann, G. M. P. (2000). *The Economics of Standardization - Final Report for Standards and Technical Regulations Directorate Department of Trade and Industry*. PhD thesis, Manchester Business School. `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8b3a594da6b780459dc0c03a9a282748d964608d`

Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable: With a new section: "On Robustness and Fragility" (Second Edition)*. Random House Publishing Group. `https://www.randomhousebooks.com/books/`

Tartaro, A. (2023). Regulating by Standards: Current Progress and Main Challenges in the Standardisation of Artificial Intelligence in Support of the AI Act. *European Journal of Privacy Law & Technologies*, 1:147–174. `https://doi.org/10.57230/EJPLT222AT`

Tartaro, A. (2024). Value-laden challenges for technical standards supporting regulation in the field of AI. *Ethics and Information Technology*, 26(4):72. `https://doi.org/10.1007/s10676-024-09809-y`

Tassey, G. (2000). Standardization in technology-based markets. *Research Policy*, 29(4):587–602. https://doi.org/10.1016/S0048-7333(99)00091-8

Teel, Z. A. (2024). Artificial Intelligence's Role in Digitally Preserving Historic Archives. *Preservation, Digital Technology & Culture*, 53(1):29–33. https://doi.org/10.1515/pdtc-2023-0050

Tessier, C. (2021). Ethique et IA : analyse et discussion. In *Conférence Nationale en Intelligence Artificielle (CNIA)*. https://hal.science/hal-03321149

Tessier, C. (2022). Parler du numérique et de son éthique: un questionnement... éthique. In *Pour une éthique du numérique*, pages 97–105. Presses Universitaires de France (PUF). https://www.puf.com/pour-une-ethique-du-numerique

Tessier, C., Chatila, R., Devillers, L., Dognin-Sauze, K., Ganascia, J., Gornet, M., and Pronesti, A. (2022). Pourquoi la reconnaissance faciale, posturale et comportementale soulève-t-elle des questionnements éthiques ? In Germain, E., Kirchner, C., and Tessier, C., editors, *Pour une éthique du numérique*. PUF.

The Hague District Court (2020). Judgment of 5 February 2020 in the matter of Nederlands Juristen Comité Voor de Mensenrechten and Federatie Nederlandse Vakbeweging versus the State of the Netherlands. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878

The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

The White House (2025). Removing Barriers to American Leadership in Artificial Intelligence. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

Thomas, R. (2018). AI Ethics Resources. *fast.ai*. https://www.fast.ai/posts/2018-09-24-ai-ethics-resources.html

Tidjon, L. N. and Khomh, F. (2022). The different faces of AI ethics across the world: a principle-implementation gap analysis. *IEEE Transactions on Artificial Intelligence*, 4(4). https://ieeexplore.ieee.org/document/9964285

Timmermans, S. and Epstein, S. (2010). A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization. *Annual Review of Sociology*, 36(1):69–89. https://doi.org/10.1146/annurev.soc.012809.102629

Toney-Wails, A., Curlee, K., and Probasco, E. (2024). Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research. In *FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2222–2233. Association for Computing Machinery. https://doi.org/10.1145/3630106.3659035

Tovo, C. (2018). Judicial review of harmonized standards: Changing the paradigms of legality and legitimacy of private rulemaking under EU law. *Common Market Law Review*, 55(4). https://doi.org/10.54648/cola2018096

Tubaro, P., Casilli, A. A., and Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society*, 7(1). SAGE Publications. https://doi.org/10.1177/2053951720919776

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59:433–460. https://www.espace-turing.fr/IMG/pdf/Computing_Machinery_and_Intelligence_A-M-_Turing.pdf

UK Parliament (2024). Artificial Intelligence (Regulation) Bill [HL]. https://bills.parliament.uk/bills/3519

Ulmer, J. T. (2015). Mesostructure. In *The Blackwell Encyclopedia of Sociology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405165518.wbeosm084.pub2

UN CEB (2024). High-level Committee on Management Task Force to Develop a System-wide Normative and Operational Framework on the use of AI in the UN System. United Nations System Chief Executives Board for Coordination. https://unsceb.org/sites/default/files/2024-02/ToR%20-HLCM%20AI%20Task%20Force%20-%20final%2011%20Jan%202024%20%281%29.pdf

UN CEB (n.d.). Inter-Agency Working Group on Artificial Intelligence. United Nations System Chief Executives Board for Coordination. https://unsceb.org/inter-agency-working-group-artificial-intelligence

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Technical Report SHS/BIO/REC-AIETHICS/2021, United Nations Educational Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Université de Montréal (2018). Montréal Declaration for responsible AI development. https://montrealdeclaration-responsibleai.com/

Vainionpää, F., Väyrynen, K., Lanamaki, A., and Bhandari, A. (2023). A Review of Challenges and Critiques of the European Artificial Intelligence Act (AIA). *International Conference on Information Systems (ICIS) 2023 Proceedings*. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/14

Vakkuri, V., Kemell, K.-K., and Abrahamsson, P. (2019). AI Ethics in Industry: A Research Framework. *arXiv preprint arXiv:1910.12695*. http://arxiv.org/abs/1910.12695

van Bekkum, M. and Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, 23(4):323–340. SAGE Publications. https://doi.org/10.1177/13882627211031257

Van Elk, K. and Van der Horst, R. (2009). Access to Standardisation. Study for the European Commission, Enterprise and Industry Directorate-General. Final Report. Technical report, EIM Business & Policy Research. https://www.anec.eu/images/Publications/Access-Study---final-report.pdf

van Leeuwen, B. (2014). PIP Breast Implants, the EU's New Approach for Goods and Market Surveillance by Notified Bodies. *European Journal of Risk Regulation*, 5(3):338–350. JSTOR. https://www.jstor.org/stable/24323461

van Leeuwen, B. (2017). *European standardisation of services and its impact on private law: paradoxes of convergence*. Number 68 in Modern studies in European law. Bloomsbury Publishing. https://www.bloomsbury.com/us/european-standardisation-of-services-and-its-impact-on-private-law-9781509930142/

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics*, 1:213–218. Springer. https://doi.org/10.1007/s43681-021-00043-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Veale, M. (2022). Value-Laden Areas for Standardisation in the AI Act. *michael veale*. https://michae.lv/value-laden-areas-in-the-ai-act/

Veale, M. and Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, (4). https://doi.org/10.9785/cri-2021-220402

Viard, T., Gornet, M., and Delarue, S. (2023a). Interpretive flexibility in data science and artificial intelligence. In *Undone Computer Science*, Nantes, France. https://undonecs.sciencesconf.org/data/Undonecs_2024_abstract_11.pdf

Viard, T., Gornet, M., and Maxwell, W. (2023b). Reading the drafts of the AI Act with a technical lens. In *Regulatable ML Workshop @ Neurips 2023*, New Orleans, USA. https://openreview.net/pdf?id=pnvRy1VzJZ

Volpato, A. and Eliantonio, M. (2020). The contradictory approach of the CJEU to the judicial review of standards: a love-hate relationship? pages 91–109. Edward Elgar Publishing. https://doi.org/10.4337/9781789902952.00010

Volpato, A. and Eliantonio, M. (2024). The participation of civil society in ETSI from the perspective of throughput legitimacy. *Innovation: The European Journal of Social Science Research*. https://doi.org/10.1080/13511610.2024.2321852

von der Leyen, U. (2019). *A Union that strives for more – My agenda for Europe – Political guidelines for the next European Commission 2019-2024*. Publications Office of the European Union. European Commission, Directorate-General for Communication. https://data.europa.eu/doi/10.2775/018127

von Ingersleben-Seip, N. (2023). Competition and cooperation in artificial intelligence standard setting: Explaining emergent patterns. *Review of Policy Research*, 40(5):781–810. https://doi.org/10.1111/ropr.12538

W3C (2024). AI & the Web: Understanding and managing the impact of Machine Learning models on the Web. https://www.w3.org/reports/ai-web-impact/

Wachter, S. (2024). Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond. *Yale Journal of Law & Technology*, 26(3). https://yjolt.org/limitations-and-loopholes-eu-ai-act-and-ai-liability-directives-what-means-european-union-united

Wachter, S., Mittelstadt, B., and Russell, C. (2021a). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review*, 123(3). https://researchrepository.wvu.edu/wvlr/vol123/iss3/4/

Wachter, S., Mittelstadt, B., and Russell, C. (2021b). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41(105567). https://doi.org/10.1016/j.clsr.2021.105567

Wagner, B. (2018). Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In Bayamlioglu, E., Baraliuc, I., Albertha, L., Janssens, W., and Hildebrandt, M., editors, *BEING PROFILED:COGITAS ERGO SUM*. Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., and Pechenizkiy, M. (2023). Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. In *FAccT'23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 805–816. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594044

Weil, P. (1983). Towards Relative Normativity in International Law? *American Journal of International Law*, 77(3):413–442. Cambridge University Press. https://doi.org/10.2307/2201073

Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research*, 74:75–109. https://doi.org/10.1613/jair.1.13196

Weizenbaum, J. (1966). ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Computational Linguistics*. https://cse.buffalo.edu/~rapaport/572/S02/weizenbaum.eliza.1966.pdf

Wentholt, I. M. E., Hoekstra, J. B. L., Zwart, A., and DeVries, J. H. (2005). Pendra goes Dutch: lessons for the CE mark in Europe. *Diabetologia*, 48(6):1055–1058. https://doi.org/10.1007/s00125-005-1754-y

Werbach, K. (2022). Orwell That Ends Well? Social Credit as Regulation for the Algorithmic Age. *U. Ill. L. Rev. 1417*. https://dx.doi.org/10.2139/ssrn.3589804

Werle, R. and Iversen, E. J. (2006). Promoting Legitimacy in Technical Standardization. *Science, Technology & Innovation Studies*, 2(1):19–39. Technische Universität Dortmund. https://doi.org/10.17877/DE290R-12756

White & Case (n.d.). AI Watch: Global regulatory tracker. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker

Whittaker, M. (2021). The steep cost of capture. *Interactions*, 28(6):50–55. https://doi.org/10.1145/3488666

Wiegmann, P. M. (2024). Voting, resilience, and inclusiveness in standardisation: a commentary on Bekkers & Lazaj's and Stanojević's implications for ETSI and beyond. *Innovation: The European Journal of Social Science Research*. https://doi.org/10.1080/13511610.2024.2321843

Wiegmann, P. M., de Vries, H. J., and Blind, K. (2017). Multi-mode standardisation: A critical review and a research agenda. *Research Policy*, 46(8):1370–1386. https://doi.org/10.1016/j.respol.2017.06.002

Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. The Riverside Press. https://monoskop.org/images/9/90/Wiener_Norbert_The_Human_Use_of_Human_Beings_1950.pdf

Wijkström, E. N. and McDaniels, D. (2013). Improving Regulatory Governance: International Standards and the WTO TBT Agreement. *Journal of World Trade*, 47(5):1013–1046. https://doi.org/10.54648/trad2013034

Williams, A. T. (2009). Taking Values Seriously: Towards a Philosophy of EU Law. *Oxford Journal of Legal Studies*, (3):549–577. https://doi.org/10.1093/ojls/gqp017

Winfield, A. (2019a). An Updated Round Up of Ethical Principles of Robotics and AI. *Alan Winfield's Web Log*. https://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

Winfield, A. (2019b). Ethical standards in robotics and AI. *Nature Electronics*, 2(2):46–48. Nature Publishing Group. https://doi.org/10.1038/s41928-019-0213-6

WTO (1995). Agreement on Technical Barriers to Trade. World Trade Organization. https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm

WTO (2000). Principles for the Development of International Standards, Guides and Recommendations. World Trade Organization. https://www.wto.org/english/tratop_e/tbt_e/principles_standards_tbt_e.htm

WTO (n.d.). Understanding the WTO: The Agreements – Standards and safety. World Trade Organization. https://www.wto.org/english/thewto_e/whatis_e/tif_e/agrm4_e.htm

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2). https://doi.org/10.1016/j.hcc.2024.100211

Yates, J. and Murphy, C. N. (2019). *Engineering Rules: Global Standard Setting since 1880*. Johns Hopkins University Press. https://muse.jhu.edu/pub/1/monograph/book/66187

Zeng, Y., Lu, E., and Huangfu, C. (2018). Linking artificial intelligence principles. https://www.linking-ai-principles.org/

Zenner, K., Marcus, J. S., and Sekut, K. (2024). A dataset on EU legislation for the digital world. Bruegel. https://www.bruegel.org/dataset/dataset-eu-legislation-digital-world

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., and Perrault, R. (2022). The AI Index 2022 Annual Report. Technical report, Stanford University Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/ai-index-report-2022/

Ziegler, W. (2020a). A Landscape Analysis of Standardisation in the Field of Artificial Intelligence. *Journal of ICT Standardization*, 8(2):151–184. River Publishers. https://ieeexplore.ieee.org/document/10258070

Ziegler, W. (2020b). A Landscape Analysis of Standardisation in the Field of Artificial Intelligence. *Journal of ICT Standardization*, 8(2):151–184. https://doi.org/10.13052/jicts2245-800X.824

Zou, J. and Schiebinger, L. (2018). AI can be sexist and racist – it's time to make it fair. https://www.nature.com/articles/d41586-018-05707-8

ECOLE
DOCTORALE

Titre : De l'IA digne de confiance aux normes techniques – l'approche particulière de l'Europe en matière de régulation de l'intelligence artificielle

Mots clés : intelligence artificielle, régulation, règlement IA, normes techniques, éthique, apprentissage machine

Résumé : L'Europe a été à l'avant-garde de l'éthique de l'intelligence artificielle (IA), en élaborant des chartes et des principes non contraignants sur l'IA ≪ digne de confiance ≫. Le terme ≪ digne de confiance ≫ est utilisé par l'Europe pour désigner les systèmes d'IA qui sont ≪ éthiques ≫, ≪ légaux ≫ et ≪ techniquement robustes ≫. L'Europe a complété ces principes non contraignants par un texte de loi sur l'IA, connu sous le nom de règlement sur l'IA, ou AI Act. Le règlement sur l'IA est l'un des premiers cadres légaux au monde à réglementer les systèmes d'IA dans différents secteurs et cas d'utilisation, en mettant l'accent sur la sécurité et la protection des droits fondamentaux. Pour les questions opérationnelles, le règlement sur l'IA s'appuie principalement sur des normes techniques en cours d'élaboration. L'approche européenne combine donc trois niveaux d'instruments réglementaires : les chartes éthiques de l'IA, le règlement sur l'IA et les normes techniques. L'approche par la normalisation est traditionnelle dans le domaine de la sécurité des produits, mais dans le règlement sur l'IA, les normes sont également censées répondre aux préoccupations en matière de droits fondamentaux. Pour éviter de faire des choix normatifs difficiles, les organismes de normalisation jouent la carte de la sécurité en élaborant des normes qui restent à un niveau élevé. De plus, dans le cadre du règlement sur l'IA, la responsabilité de l'élaboration des normes techniques est déléguée à des organismes de normalisation privés, où les grandes entreprises multinationales sont surreprésentées et exercent une influence considérable. Ces normes sont également généralement payantes, bien que la situation puisse évoluer dans les années à venir après une récente jurisprudence de la Cour de justice de l'Union européenne. Les experts en normalisation sont donc sous pression pour fournir des normes à temps et de bonne qualité.

Title : From trustworthy AI to technical standards – The distinctive European approach to artificial intelligence regulation

Keywords : artificial intelligence, regulation, AI Act, technical standards, ethics, machine learning

Abstract : Europe has been at the forefront of Artificial Intelligence (AI) ethics, developing non-binding charters and principles on "trustworthy" AI. The term "trustworthiness" is used by Europe to designate AI systems that are "ethical", "legal" and "technically robust". Europe has supplemented these non-binding principles with a binding regulation on AI, known as the AI Act. The AI Act is one of the world's first comprehensive frameworks for regulating AI systems across different industries and use cases, focusing on safety and protection of fundamental rights. The AI Act relies, for operational questions, mostly on technical standards that are in the course of development. The European approach thus combines three layers of regulatory instruments: AI ethics charters, the AI Act and technical standards. The standardisation approach is traditional in product safety, but under the AI Act, standards are also expected to address fundamental rights concerns. To avoid making hard normative choices, standardisation organisations are playing it safe, developing standards which remain at a high-level. Moreover, under the AI Act, the responsibility for developing technical standards is delegated to private standardisation bodies, where large multinational companies are over-represented and hold significant influence. These standards are also often locked behind paywalls, although the situation may evolve in the coming years after a recent case law from the Court of Justice of the European Union. Standardisation experts therefore face pressures to deliver standards on time and of good quality.

Institut Polytechnique de Paris
91120 Palaiseau, France