Global AI policy in the making: actors, framings, and approaches

Author: Lewin Schmitt

TESI DOCTORAL UPF
2024

THESIS SUPERVISOR: Prof. Dr. Jacint Jordana Casajuana
Department of Political and Social Sciences

upf. Universitat Pompeu Fabra Barcelona

«Freiheit ist immer die Freiheit der Andersdenkenden.»

Rosa Luxemburg

Acknowledgements

This thesis is all about artificial intelligence, but it would be nothing without the human intelligence, compassion, and support that I have received from so many throughout this journey. I am grateful for each encounter and exchange that helped me grow personally and academically. While the following enumeration is clearly incomplete, I would like to acknowledge at least some of the people who defined these formative years.

First of all, I thank Jacint, my supervisor, for making this entire project possible in the first place. These were some of the happiest years of my life, but the PhD also confronted me with some of the most challenging moments. Jacint steered me through it all with the right mix of patience, guidance, trust, and loyalty.

He also assembled a brilliant doctoral examination board and I am indebted to the reviewers for their insightful comments and suggestions, which profoundly improved the quality of my work. The board members committing their valuable time and energy to reading this thesis, reviewing it, and also attending the defence truly honours and humbles me. All remaining errors are mine and mine alone.

I would also like to acknowledge all other reviewers – anonymous or not – as well as discussants, colleagues, and students, who have engaged with my work throughout the years and, in some way or another, shaped and sharpened my understanding of the world.

Then, there is Adam, who was forced to share with me the most beautiful office of IBEI (facing the sunniest rooftop of Barcelona). As if that was not enough, we also shared beers, gossip, and a predilection for fine electronic music that kept us focused during our work on the countless GLOBE project deliverables.

There are my colleagues and friends from IBEI and the UPF PhD programme who formed a community of comradeship. There are too many to name them all, but special shout-outs go out to Carlos Bravo, who made me feel welcome and oversaw my initiation; to Guillem, eternal discussant and fellow R nerd; to Diego, co-author and recurring host of unforgettable movie nights; to Shashwat, sporadic host of unforgettable election nights and arguably the best Indian Flammkuchen chef in town; to Erick, my dear lunch buddy and confidant; to Jana, Abe, and Leo for those heartwarming family vibes; to Hana, for her refreshing yet soothing vibe (and the occasional cigarette); to Laia, for not getting mad at me when I failed to deliver on our project; and to Aitor, el juergas (o el fiestas?), trusted co-author and compañero. A big thank you also to Carlos Sanchez, Ana Salas, and the remaining staff at IBEI and UPF whose work may often be invisible but definitely is essential.

Furthermore, I am grateful to my circle of friends in Barcelona and beyond, who kept me sane and well-guarded from any aspirations to hide in the academic ivory tower or the metaphorical prison of the university library. Most importantly, to José and Mar, who convinced me that Barna was the place to be and on whose wonderful friendship I could always count; to the entire Latinx gang, whose fiestas and bailes kept me young and wild and free; to the beach volleyball community, which kept me fit, tanned, and level-headed; to my flatmates, which helped me feel at home so far from home; and to all the incredible friends from previous chapters of my journey who enrich my life in so many ways.

Then, of course, to my family and – above all – my parents, for their unconditional love, support and encouragement, for the liberties and opportunities which they grant me. And to my dear, dear brother, who inspires and amazes me every single day.

Finally, to Jana, my partner in crime who single-handedly pushed me over the finish line with her sharp editorial eye and tireless motivating interventions. I cannot thank you enough for your attentive assistance and welcomed distractions, for all the moral and technical support, and for your patience and understanding whenever doubts, stress, or frustration got the better of me. I owe you one!

Abstract

The rapid advances in artificial intelligence (AI) technologies have sparked regulatory and political activities in jurisdictions around the world. This dissertation outlines and investigates the nascent domain of global AI policy, using a variety of methods and presenting novel empirical data to substantiate the discussion. The three articles predominantly focus on actors and their preferences for different approaches to AI governance. Through qualitative and quantitative document analyses, this research illustrates how stakeholders frame the AI policy debate according to their interests. For one, the findings reveal international organisations' high levels of agency in addressing global AI policy and a tendency to address new challenges within existing governance frameworks. Moreover, the analysis of national AI strategies shows that democratic governments have blind spots when it comes to the impact of AI on democracy. Lastly, the dissertation uncovers systematic differences across sectors and regions regarding the salient innovation-protection trade-off. But it also identifies a certain alignment between the American and European public sectors, which might be indicative of closer international cooperation in the future.

Resumen

Los avances de las tecnologías de inteligencia artificial (IA) han espoleado un frenesí de actividades normativas, regulatorias y políticas en jurisdicciones de todo el mundo. En esta tesis se describen e investigan las actividades y los resultados de la política mundial de IA, utilizando diversos métodos y presentando datos empíricos novedosos para fundamentar el debate. Los tres artículos se centran principalmente en los actores y sus diferentes preferencias por determinados enfoques de la gobernanza de la IA. Mediante el análisis cualitativo y cuantitativo de documentos, esta investigación ilustra cómo las partes interesadas enmarcan el debate político sobre la IA. Por un lado, los resultados revelan el alto grado de agencia de las organizaciones internacionales a la hora de abordar la política mundial de IA y su tendencia a abordar los nuevos retos dentro de los marcos de gobernanza existentes. Por otra parte, el análisis de las estrategias nacionales de IA muestra que los gobiernos democráticos tienen puntos ciegos cuando se trata del impacto de la IA en la democracia. Por último, el tercer artículo descubre diferencias sistemáticas entre sectores y regiones. Pero también identifica una cierta alineación entre el sector público estadounidense y el europeo, que podría ser indicativa de una cooperación internacional más estrecha en el futuro.

Resum

Els ràpids avenços de les tecnologies d'intel·ligència artificial (IA) han estimulat un frenesí d'activitats polítiques i normatives a les jurisdiccions d'arreu del món. Aquesta tesi descriu i investiga les activitats i els resultats de la política global d'IA, utilitzant una varietat de mètodes, i presenta dades empíriques noves per corroborar la discussió. Els tres articles es centren principalment en els actors i les seves diferents preferències per a determinats enfocaments de govern de la IA. Mitjançant una anàlisis de documents qualitatius i quantitatius, aquesta investigació il·lustra com les parts interessades enmarquen el debat sobre la política d'IA. D'una banda, les troballes revelen els alts nivells d'agència de les organitzacions internacionals per abordar la política global d'IA i una tendència a abordar nous reptes dins dels marcs de govern existents. A més, l'anàlisi de les estratègies nacionals d'IA mostra que els governs democràtics tenen punts cecs quan es tracta de l'impacte de la IA en la democràcia. Finalment, el tercer article descobreix diferències sistemàtiques entre sectors i regions. Però també identifica una certa alineació entre el sector públic nord-americà i europeu, que podria indicar una cooperació internacional més estreta en el futur.

Contents

Introduction

Over the past years, artificial intelligence (AI) technologies have developed at a remarkable pace, transforming business practices and the digital sphere, and elevating AI from a niche topic for sci-fi nerds and computer scientists to a seemingly ubiquitous concern of public debate. In 2024, the interest in AI is enormous. The public imagination is captured by impressive advances in AI-powered text and media generation, alongside abstract reasoning capabilities that often rival or even surpass human capacities. This has given additional urgency to policymakers trying to mitigate the technology's risks while harnessing its opportunities. The European Union (EU) has finalised negotiations for a groundbreaking horizontal legislation, the AI Act; China has recently passed a number of specific rules and obligations for AI developers; and even the United States, which traditionally followed a hands-off approach leaving most of Big Tech's activities unchecked, is seriously considering stricter policy responses. In the midst of this global momentum for technology regulation, other jurisdictions around the world are following suit. Undoubtedly, AI is now not only a regular conversation topic at dinner parties but also a priority for policymakers and regulators. And rightly so: the transformative impact of AI is already felt by many, but by all accounts will accelerate even further in the coming years.

So what has happened in the last years that explains global AI policy activities and outcomes? Of course, technological progress plays a paramount role as a catalyst triggering regulatory activities. Yet, these activities could plausibly have taken multiple different routes, begging the question of why particular pathways have unfolded. The three articles of this dissertation address various aspects related to this overarching question. They identify key actors, events, and approaches. In this, they systematically analyse different stakeholders' framings and preferences in the global AI policy debate. Anchored in the perspective of global governance, the analysis is not restricted to nation-states, but focuses heavily on international organisations and fora, as well as non-state actors. Below, I sketch out a synopsis of the dissertation's main research questions and contributions, before providing a more extensive conceptual discussion in which the three articles are embedded.

In brief, the first article ("Mapping global AI governance: a nascent regime in a fragmented landscape", chapter 1) asks about the nature and trajectory of the nascent global AI governance regime, and whether we can see signs of consolidation amongst the overall fragmentation of approaches. It describes and evaluates the emerging global AI governance architecture, identifying the key actors and initiatives in the international arena, which all gravitate around the Organisation for Economic Co-Operation and Development (OECD). As its analytical framework, the article differentiates between state- and non-state-led initiatives, positioning them within or outside the current global governance architecture. It thus highlights international organisations' significant role in addressing global AI policy and a tendency to tackle emerging challenges within established frameworks.

The article was published in the journal *AI Ethics* in 2021 and has since been cited in multiple studies. As highlighted in the article, AI is a fast-developing field, and the relevant policy and governance fora and tools are only beginning to emerge. Accordingly, some aspects of the article might already require updates. Importantly, however, the key assertions hold true, as evidenced by later developments and additional research. Most notably, the OECD has maintained its key role within the global governance regime. This has been demonstrated, for instance, by outcomes of the G7's Hiroshima AI Process, which has been carried out in close cooperation with the OECD. In October 2023, the group's leaders agreed on international guiding principles on AI and a voluntary Code of Conduct for AI developers. Notably, the principles explicitly refer to the OECD's 2019 AI Principles as the foundation, stressing their ongoing relevance.

The second article ("AI and democracy: A blind spot in national AI strategies?", chapter 2) is interested in the way that governments address AI's impact on democracy. It first provides a novel conceptual framework to organise the various ways in which AI developments can be harmful or beneficial to the health of democratic systems. Second, the article applies this approach to conduct a qualitative in-depth analysis of 29 OECD member states' national AI strategies. The analysis unveils significant blind spots, attributed to the prevalence of technocratic approaches that emphasise economic opportunities and potential efficiency gains in the public sector. Moreover, it unearths considerable variation in both the degree and the concrete frames with which governments address the different dimensions of democracy. The most widespread cause for concern relates to AI's impact on civil liberties, but many strategies also show awareness of the possible dangers to participation, equality, and accountability.

Lastly, the third article ("Transatlantic perspectives on AI policy: shifts in public sector preferences", chapter 3) presents an original data set of over 300 AI policy documents, one of the biggest collections to date. It employs novel computational text analysis methods to investigate the different preferences of AI regulation by various stakeholders, asking whether we can observe meaningful differences across the public sector, businesses, and civil society, and across regions. Demonstrating the existence of such differences across both sides of the Atlantic, it enriches our understanding of

stakeholders' competing preferences and framings. It also contextualises recent shifts in how the American and European public sector have approached AI regulation, which demonstrates a certain convergence between the two. Whereas debates in Europe are turning towards more innovation-friendly positions in a reaction to the block's initially highly cautionary approach, the US is on the opposite course. After an initial laissez-faire, pro-business stance, the American public sector is now increasingly attuned to the associated risks and the need for some regulation and protection against potential harms stemming from AI, such as algorithmic discrimination, privacy violations, or faulty outputs and security flaws.

Research design and methods

The three articles of the thesis draw on a broad range of methods from qualitative and quantitative research traditions, depending on the research question and available data. Overall, research for this dissertation has mainly relied on desk research and the analysis of primary and secondary data. Throughout the process, my research has been informed by numerous background talks and exchanges with relevant stakeholders. Empirical evidence is largely drawn directly from document sources, which I processed statistically through qualitative and quantitative, computational text analysis methods.

The first article is mostly interpretative, tracing the emergence of the global AI governance regime based on primary and secondary sources, from media reporting to stakeholders' public statements. In addition, by distinguishing between state-led and non-state-led governance initiatives, it presents an important conceptual contribution to the then-nascent literature on global AI governance.

The second article presents a qualitative content analysis of national AI strategies. It thus draws directly on primary sources, namely 29 documents resembling governments' official AI strategies. Descriptive statistics and inference are used to derive high-level findings from the analysis and to quantitatively support the drawing of some overarching conclusions. Conceptually, the paper offers a novel analytical framework which allows classifying the myriad ways in which AI may impact democracy.

Lastly, the third article explores more innovative text-as-data methods, using computational text analysis and zero-shot predictions by a large language model (GPT3.5-turbo). Moreover, it presents a large-scale data collection effort, gathering over 300 AI policy documents in a machine-readable format. The operationalisation of the dependent variables, which act as proxies for stakeholders' preferences on AI regulation, presents another methodological contribution. The subsequent quantitative analysis provides substantiation for the assumptions presented in the article, while also showing the limitations of the approach. Taken together, it thus paints a rich and detailed picture of how actors from different key sectors frame AI ethics and policy issues.

The combination of various methodological approaches and different empirical sources underlines the value of mixed-methods research designs in political science. The use of an advanced AI tool is not only an original meta-commentary on the subject of this thesis, but also provides a cutting-edge exploration and application of a new instrument in the empirical social scientist's toolbox. Overall, the dissertation's three articles shed light on various important facets related to the emergence of global AI policy. From stakeholder mapping to qualitative and quantitative document analysis, they put forward comprehensive interpretative and descriptive analyses. Making use of a mixture of quantitative and qualitative methods, building on a range of sources, and providing both conceptual and empirical contributions, the dissertation offers a rich and informative account of global AI policy in the making.

The remainder of this introduction shall provide the reader with the necessary definitions and key concepts, and situate them within the broader literatures to which the articles speak. Lastly, I will also widen the scope and reflect on some broader developments, which form the backdrop of this research.

Defining artificial intelligence

As a general-purpose technology, AI escapes simplistic definitions. Broadly speaking, the term is meant to capture computer software and statistical modelling techniques that aim to replicate human intelligence. While certain contexts may require carefully distinguishing between the various strands of machine learning and approaches, such as neural networks or symbolic reasoning, this dissertation deploys a more comprehensive conception of AI that generously encapsulates its many facets. This is in line with most of the AI policy debate, which is generally more concerned with the consequences or outcomes of technology, and less so with the exact definitions or technological functionality. Consider AI used in informing a bank's credit loan decisions: For most stakeholders – the bank, its consumers, or the regulator – it does not matter whether the AI model is built on a random forest or regression technique. What matters is whether the automated decisions that the bank derives from the model output are accurate, fair, and explainable. The finance sector is a prominent example where AI is already being deployed heavily. Other sectors with a high AI absorption rate include health, agriculture, and of course software/digital services. While it is important to take into account sector-specific nuances of AI policy, I purposefully approach the topic with a universal, sector-overarching lens. This is again in line with a substantial part of the relevant policy debate and regulatory activity.

AI in its current form is an emerging technology according to the criteria by Rotolo, Hicks, and Martin (2015): recent advances represent radical novelty, fast growth, coherence, prominent impact, uncertainty, and ambiguity. Critically, while few observers can agree on the exact trajectory that AI developments will take, few have doubts about its ongoing acceleration and transformative impact. This matches the definition of emerg-

ing technologies, whose "most prominent impact, however, lies in the future and so in the emergence phase is still somewhat uncertain and ambiguous" (ibid.). This is usually coupled with gaps in adequate governance mechanisms, policies, or regulation – or even the total absence thereof. It is precisely the filling of these gaps that "global AI policy in the making" aims to describe, illustrate, and discuss.

Global governance, global policy, and regulation of AI

As this dissertation approaches the nexus of AI and global policy, it situates the studies within a broader scholarly discourse on global governance, global policy, and international regulatory regimes. Drawing upon various insights from these literatures, it aims to provide a comprehensive understanding of the complexities inherent in the intersection of AI and global affairs.

The focus on the international level is intentional. By definition, as a transformative technology with strong cross-border implications, AI demands an inherently global perspective, as national discussions and policy trajectories can only be properly accounted for by looking at the international level, and vice-versa.

So what do I mean precisely when I speak of global governance, policy, and regulation in the context of AI? First of all, global AI policy includes all formal and informal policy activities that guide the development, deployment, and use of AI technologies on a global scale. By global scale, I mean that they go beyond national borders, either directly and by design (e.g., international agreements, extraterritorial legislation, ...) or indirectly and at times even unintentionally (e.g., standards and policy diffusion). As such, global AI policy encompasses the guidelines and frameworks set by various actors, including governments, international organisations, and industry bodies, to address the ethical, legal, and societal challenges posed by AI. In its broadest sense, global AI policy debates are then all those interventions by relevant stakeholders that shape or intend to shape global policy on AI.

AI regulation, in turn, refers to specific rules, laws, or legal frameworks enacted by governments (or, in the case of the EU, supranational jurisdictions) to govern and control the use of AI technologies. Compared to broader policies, regulation is more focused and prescriptive, legally binding and enforceable. This may include requirements for transparency, accountability, data protection, and compliance with certain standards.

Global AI governance refers to the broader and more holistic management of AI's impact on society, including not only policy-making but also the coordination of efforts to ensure responsible, transparent, and beneficial AI advancements across borders. It spans a wide range of issues, such as the publication of ethics guidelines for AI development (Jobin, Ienca, and Vayena 2019), hypothetical international agreements to prevent a superpower arms race (Dafoe 2018), or the design of global institutions and governance regimes (Cihon, Maas, and Kemp 2020; Stix and Maas 2021). So while global

AI policy focuses on specific rules and guidelines, global AI governance takes a comprehensive approach by looking also at soft law and non-legal, non-policy initiatives such as cross-border collaboration, sharing best practices, and addressing issues beyond policy-making. Notably, the two concepts are intrinsically related. Effective global AI governance often relies on well-structured policies, while policy-making benefits from a broader governance framework to address the multifaceted implications of AI.

Out of this, I derive a conceptual hierarchy which also structures the setup of the thesis: *governance* > *policy* > *regulation*. The first article of this dissertation (1) maps the emerging global governance regime of AI, identifying key stakeholders and activities and offering an analytically-based interpretation of its defining characteristics and likely future developments. The second article (2) is then more concerned with national strategies as expressions of various AI policy approaches. Lastly, the third article (3) – while studying the broader global AI policy debate – is interested in identifying preferences relevant to AI regulation, the most concrete of the three levels.

It is worth pointing out that *governance*, especially in the context of AI, is a diffuse concept with multiple interpretations. In a technical context, it can also refer to the rules and procedures steering an AI product (e.g., OpenAI instructing its chat bots not to respond to harmful or unethical requests). In a corporate context, it can refer to the structures and institutions within or across companies that are designed to steer the development of AI (e.g., supervision boards, an AI ethics officer, ...). In the AI ethics and policy context that is most relevant to this dissertation, governance refers to the broad set of formal and informal rules and processes – ranging from legislation and regulation to industry standards and self-binding commitments – that together define the operational boundaries for AI developers and users.

Erman and Furendal (2022) offer a conception in which the study of AI governance is but a subdomain of AI ethics. For this thesis, the conceptual relationship is reversed, with governance being the umbrella that spans ethics and policy debates as well as regulation. Global AI governance, by extension, refers to those aspects of AI governance that have a clearly international element. As such, it is part of the global AI policy environment, though as mentioned above, it goes beyond mere policy activities. Global AI governance is derived from the concept of global governance, a framework that acknowledges the increasingly interconnected and interdependent nature of global challenges, and the importance of considering a diverse set of actors (Weiss 2000). It is thus better suited for the study of global AI policy than traditional conceptions of governance, which are often rooted in state-centric models and face significant challenges when applied to the inherently transnational nature of AI technologies. Scholars such as Rosenau (1992) and Held (1996) argue that global governance involves a networked approach, where state and non-state actors collaboratively address issues transcending national boundaries. The literature on global regulatory regimes also informs the dissertation's conceptual framework. Keohane and Victor (2011) emphasise the importance of understanding the preferences and strategies of diverse actors in shaping global reg-

ulations. Contemporary conceptions of global policy are similarly attuned to the important role of non-state actors. Stone (2019) describes global policy as incorporating "both governmentally steered processes of 'international public policy', better known today as 'trans-governmentalism', and 'transnational policy processes' where there is a greater degree of authoritative steering from non-state actors." To account for this, I will elaborate on the role of actors in the following section.

Stone (ibid.) also presents the concept of 'epistocracy' as an essential power dynamic behind global policy making. The concept aims "to capture knowledge-based decision-making as well as knowledge networking between states and in transnational policy communities", and serves to highlight how these dynamics concentrate "political power among those with superior knowledge of the complexity of public problems and policy processes." The findings of my first paper, which identify the OECD as the backbone of an epistemic authority, speak to this notion of epistocracy and the power of knowledge in shaping policies for emerging technologies such as AI.

Actors in global AI policy

The global governance of AI is characterised by a multi-stakeholder paradigm, wherein states, intergovernmental organisations, non-governmental entities, and private industry collaborate and compete in defining norms, standards, and regulations (Cihon, Maas, and Kemp 2020; Veale, Matus, and Gorwa 2023). According to global governance perspectives, understanding the intricate web of international stakeholders becomes crucial, as these actors collectively shape the governance structures and policies for this emerging technology. Indeed, as the thesis title highlights, my dissertation emphasises the role of – institutional – actors in the development of global AI policy. Accordingly, by mapping international stakeholders in global AI governance, the first paper contributes to this understanding by illustrating the diverse array of actors involved in shaping the rules and norms governing AI technologies.

One set of actors, which has an elevated role in shaping AI policy, contains national governments. AI policy is by and large a matter dealt with on the national level (though there are numerous and important exceptions to both the sub- and supranational level). Therefore, the second article analyses national AI strategies as one prominent feature of AI policy developments.

Another important consideration related to actors is the question of agency. On the one hand, actors of global AI policy can be seen as subjects, actively engaged in policy discussions and policy-making; on the other hand, they can be seen as objects that are affected by the outcomes of these policy processes. The second and third paper speak to this dual mechanism. In addition, there is a similar bi-directional mechanism at play which lies at the heart of the second paper: while policy, governance, and regulation are important factors shaping the development of AI technologies, these technologies in turn can also affect political processes (Boix 2022; Gilardi 2022).

Earlier, I mentioned that this dissertation is not restricted to nation-states, but focuses heavily on intergovernmental and multilateral actors as well as domestic and international non-state actors, such as business and civil society organisations. Indeed, my first article shows that global AI policy debates are marked by the engagement of a multitude of international organisations, both state-led and non-state-led. Complementary research has equally identified the importance of different sets of stakeholders, including the public and business sectors, and civil society (see also Jobin, Ienca, and Vayena 2019; Fjeld et al. 2020; Hagendorff 2020; Schiff et al. 2021; Ulnicane et al. 2022). These groups of actors have different preferences when it comes to AI policy and regulation, which is the focus of my third article.

Moreover, they have different means for influencing policy processes and trying to align outcomes with their preferences (Tallberg, Lundgren, and Geith 2023). Further below, I discuss how they engage in the policy debate through the publication of AI principles and policy documents. While this is an observable and relatively transparent approach, actors can also retort to other means, such as lobbying and informal networking. And on the intergovernmental level, many important negotiations are going on behind closed doors. Naturally, all these activities are harder to study from an outside perspective. Nevertheless, future research should also look for ways to shine light on these covert or obscure actions to further enrich our understanding of how AI policy is made.

This brief discussion has also demonstrated the absence of a crucial group of stakeholders from global AI policy discussions: citizens. Indeed, they are rarely involved in these processes (Schiff 2022), and generally have to hope that their interests are represented by civil society organisations or their governments. While much of this is normal in international negotiations and global policy-making (national and sub-national AI policy developments are indeed often more attuned to the involvement of citizens), it is nevertheless problematic since it excludes valuable critical perspectives. Participatory processes around global climate change governance and the UN's Sustainable Development Goals have demonstrated that it is possible to include citizens even in highly complex international policy discussions (Worthington, Rask, and Minna 2013).

Another set of stakeholders that is largely absent from the global AI policy debate – or at least severely underrepresented – are voices from the Global South and from marginalised communities (Hickok 2021; Roche, Wall, and Lewis 2023). I will return to this further below.

Different approaches to global AI policy

Shaped by their respective interests and distinct roles, different actors pursue and prefer different approaches to global AI policy. This thesis aims to highlight some of the differences while also showing the commonalities on which successful global governance of AI can be built.

Regarding the design of the wider global AI governance regime, approaches can either turn to creating new institutions and mechanisms, or try to accommodate the new challenges brought about by AI through existing avenues. The first article speaks to this in its conceptual and analytical contributions, identifying a tendency to address new challenges within existing frameworks.

Moreover, approaches to AI policy can differ when it comes to the issues that are prioritised or addressed at all. Actors can give more or less weight to different aspects and thus shape approaches accordingly. In this light, the second and third articles investigate how different actors frame issues around AI, and what we can learn from studying these framings.

In addition, the way that a political actor frames a certain issue reveals important information about their preferences, concerns, and priorities. This is why the second and third articles are concerned with investigating actors' framings of AI policy and regulatory issues. Understanding different actors' preferences and framings matters because they play an influential role in the development of new regulations on AI (Ulnicane et al. 2021; Tallberg, Lundgren, and Geith 2023). By shaping the agenda-setting and problem-definition stage of the policy cycle, they can condition certain policy outcomes (Elder and Cobb 1984; Gilardi, Shipan, and Wüest 2021). Malmborg (2023) showed the important role that narratives played internally and externally in the European Commission's policy-making process. Moreover, AI policy documents may serve as cue givers for policymakers, perform signalling functions to stakeholders, and influence media coverage and thus, ultimately, public opinion.

AI and democracy

There are many ways in which the growing presence of AI systems may affect society and individuals, for better or worse. Topics that have received significant attention from scholars and policymakers include AI's impact on labour markets, algorithmic biases, and the potential for innovative solutions to improve processes and outcomes across use cases in diverse sectors such as agriculture, health, or customer services. An area that has received comparably less attention is the nexus of AI and democracy (but see Manheim and Kaplan 2019; Boix 2022; Djeffal 2022; Coeckelbergh 2024).

As the previous sections and the articles of this dissertation argue, policy is crucial in steering technological developments. But of course, the reverse relationship is just as important. With AI advancing and becoming ever more ubiquitous, researchers,

policymakers, and society at large need to inquire how the technology and its use affect our political systems, institutions, and processes.

So-called deepfakes – AI-generated audio or video impersonations – have targeted politicians in multiple instances across the world. In late 2023, an advocacy group in Germany has made headlines with a doctored video showing Chancellor Olaf Scholz apparently announcing that his government will outlaw the far-right opposition party Alternative for Germany or AfD (Dörner 2023). Deepfake material also emerged just days before the 2023 elections in Slovakia. Unauthentic audio recordings that supposedly implicated leading candidates in an election fraud scheme were released hours before the election, leaving just enough time for the content to go viral, but not enough for it to be widely debunked as fake news (Meaker 2023). In the subsequent election, the progressive camp of the affected candidates narrowly lost. But there are also less malign instances of the use of AI in democratic contexts. For instance, the mayor of New York City, Eric Adams, has used AI to make robocalls to citizens in languages he does not actually speak (Calma 2023). Such misleading practices, even if well-intended, risk polluting the information environment and pose serious challenges to established democratic norms and practices.

These episodes illustrate the need to study carefully the various ways in which AI will – and to some extent already does – impact democracy, for better or worse (Nemitz 2018; Helbing et al. 2018; Zarkadakēs 2020; Boix 2022; Acemoğlu 2021). To organise these effects systematically, the second article puts forward a novel conceptual framework that categorises AI's impact on the main dimensions of democracy, covering participation, control, and performance. It discusses the plausible ways in which various facets may be affected such that they enhance or undermine democratic norms and prerequisites, showcasing the main levers through which such developments may occur. I then proceed to apply this framework to a qualitative, in-depth review of 29 OECD governments' national AI strategies, asking critically whether they are sufficiently alert to the risks and sensitive to the opportunities. As the analysis shows, there are huge blind spots when it comes to governments' attention to many of these issues. Democratic governments predominantly revert to technocratic approaches, framing AI as a technical, competitive challenge with economic ramifications, and rarely engaging with the more sensitive – and necessary – political questions, such as the future of democracy.

The big picture

The various elements of this dissertation also feed into broader debates regarding AI ethics, global AI policy, and the geopolitics of AI. First of all, the examination of business sector interests speaks to contributions that critically engage with the elevated role of corporate actors in driving technology policy, and especially AI policy. As Veale, Matus, and Gorwa (2023) point out, the global AI landscape is characterised by very high levels of industry control and influence, rendering it important to carefully consider and recalibrate the degree of business power on debates that may well shape humanity's future trajectory. The outsized role of the private sector primarily stems from the importance that US Big Tech has in the development of AI. In light of this, it should be concerning that democratic governments seem reluctant to engage with profound political questions in their AI strategies and that only few of them emphasise the role of civil society and citizen involvement in these processes.

Second, we are entering a crucial phase in the development of global and national AI policies, which lets us observe the key role of framings and narratives in shaping policy outcomes. A clear and timely illustration of this can be found in the negotiations around the EU's AI Act: it pitches innovation- and business-friendly positions against precautionary, sceptical voices (Justo-Hanani 2022; Tallberg, Lundgren, and Geith 2023). While the former are generally embracing AI as a harbourer of economic growth and other improvements, the latter are concerned about the technology's impact on fundamental rights such as privacy and non-discrimination, and about humanity's long-term ability to maintain autonomy over AI. Each camp has opposing views on how AI policy should look, with one proposing a laissez-faire approach, while the other wants to regulate or even ban some use cases of AI, such as facial recognition in public places. Thus, policy outcomes (i.e., the resulting national and international policies and regulations of AI) are key for determining future developments and impacts of AI.

The struggle between different visions of AI can also be observed at an even larger scale: globally, the USA are competing with China in a race for AI supremacy, with the EU aiming to establish itself as a third player (Cath et al. 2018; Roberts, Cowls, Hine, Mazzi, et al. 2021). In this race, the USA is generally following a pro-business approach with little regulatory intervention (Robles and Mallinson 2023). The Chinese government, on the other hand, has put strong obligations on its AI companies, but simultaneously pursues a strategy that relies on public sector involvement and tight state control (Roberts, Cowls, Hine, Morley, et al. 2021). In both cases, governments are heavily investing in AI RDI, conceiving of the technology as a prerequisite to dominate world affairs in the 21st century. The EU, in its third-way approach, aims to establish itself as a leader in what it frames as "trustworthy AI", emphasising ethics and responsible design (Justo-Hanani 2022). Only time will tell which approach is most successful. But this dissertation and a growing literature on AI policy highlight that the political choices we

make today will have profound impacts tomorrow, given AI's transformative potential.

The schematic account of global AI governance as a race between China, the USA, and the EU reveals another important realisation: voices from the Global South are largely absent from the debate (see Hickok 2021; Roche, Wall, and Lewis 2023). This is highly problematic, since the impacts will be felt just as acutely – or even more so – in poorer societies. Workers across developing countries are either at risk of having their jobs replaced by an AI, or already labour for AI companies – often under problematic conditions (Altenried 2020; Perrigo 2023; Tan and Cabato 2023). Both my first and third articles found clear indications of the underrepresentation of stakeholders from the Global South. While the focus on the OECD in the second article was purposeful, it is worth noting that the review of government strategies found barely any consideration of questions around global inequalities, indicating a grave neglect.

These are some aspects that underline the importance of grappling with the actors, framings, and approaches of global AI policy as it is in the making. It is my hope that the reader of this dissertation finds merit in the following articles, which aim not only to contribute to the academic scholarship but also inform evidence-based policy-making. As we stand at the gates of the AI age, we must carefully weigh our choices and forcefully pursue the path that allows humanity to harness the many opportunities while mitigating the various risks.

Bibliography

Acemoğlu, Daron (2021). *Dangers of unregulated artificial intelligence* (cit. on p. 10).

Altenried, Moritz (2020). "The platform as factory: Crowdwork and the hidden labour behind artificial intelligence". In: *Capital & Class* 44.2. Publisher: SAGE Publications Ltd, pp. 145–158. DOI: 10.1177/0309816819899410 (cit. on p. 12).

Boix, Carles (2022). "AI and the Economic and Informational Foundations of Democracy". In: *The Oxford Handbook of AI Governance*. Ed. by Justin B. Bullock et al. 1st ed. Oxford University Press. DOI: 10.1093/oxfordhb/9780197579329.013.64 (cit. on pp. 7, 9, 10).

Calma, Justine (2023). "NYC Mayor Eric Adams uses AI to make robocalls in languages he doesn't speak". In: *The Verge* (cit. on p. 10).

Cath, Corinne et al. (2018). "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach". In: *Science and Engineering Ethics* 24.2, pp. 505–528. DOI: 10.1007/s11948-017-9901-7 (cit. on p. 11).

Cihon, Peter, Matthijs M. Maas, and Luke Kemp (2020). "Fragmentation and the Future: Investigating Architectures for International AI Governance". In: *Global Policy* 11.5, pp. 545–556. DOI: 10.1111/1758-5899.12890 (cit. on pp. 5, 7).

Coeckelbergh, Mark (2024). *Why AI undermines democracy and what to do about it*. Cambridge Hoboken, NJ: Polity (cit. on p. 9).

Dafoe, Allan (2018). *AI Governance: A Research Agenda* (cit. on p. 5).

Djeffal, Christian (2022). "Democracy, AI Regulation and the Draft EU AI Act". In: *Turkish Policy Quarterly* (cit. on p. 9).

Dörner, Jan (2023). "Deepfake-Video von Kanzler Scholz sorgt für Empörung". In: *Berliner Morgenpost* (cit. on p. 10).

Elder, Charles D. and Roger W. Cobb (1984). "Agenda-Building and the Politics of Aging". In: *Policy Studies Journal* 13.1, pp. 115–129. DOI: 10.1111/j.1541-0072.1984.tb01704.x (cit. on p. 9).

Erman, Eva and Markus Furendal (2022). "Artificial Intelligence and the Political Legitimacy of Global Governance". In: *Political Studies* (cit. on p. 6).

Fjeld, Jessica et al. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society (cit. on p. 8).

Gilardi, Fabrizio (2022). *Digital Technology, Politics, and Policy-Making*. 1st ed. Cambridge University Press. DOI: 10.1017/9781108887304 (cit. on p. 7).
Gilardi, Fabrizio, Charles R. Shipan, and Bruno Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". In: *American Journal of Political Science* 65.1, pp. 21–35. DOI: 10.1111/ajps.12521 (cit. on p. 9).
Hagendorff, Thilo (2020). "The Ethics of AI Ethics: An Evaluation of Guidelines". In: *Minds and Machines*. DOI: 10.1007/s11023-020-09517-8 (cit. on p. 8).
Helbing, Dirk et al. (2018). "Will Democracy Survive Big Data and Artificial Intelligence?" In: *Towards Digital Enlightenment*. Ed. by Dirk Helbing. Cham: Springer International Publishing, pp. 73–98. DOI: 10.1007/978-3-319-90869-4_7 (cit. on p. 10).
Held, David (1996). *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Stanford: Stanford University Press (cit. on p. 6).
Hickok, Merve (2021). "Lessons learned from AI ethics principles for future actions". In: *AI and Ethics* 1.1, pp. 41–47. DOI: 10.1007/s43681-020-00008-1 (cit. on pp. 8, 12).
Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9. ZSCC: 0000007, pp. 389–399. DOI: 10.1038/s42256-019-0088-2 (cit. on pp. 5, 8).
Justo-Hanani, Ronit (2022). "The politics of Artificial Intelligence regulation and governance reform in the European Union". In: *Policy Sciences* 55.1, pp. 137–159. DOI: 10.1007/s11077-022-09452-8 (cit. on p. 11).
Keohane, Robert O. and David G. Victor (2011). "The Regime Complex for Climate Change". In: *Perspectives on Politics* 9.1. ZSCC: 0001434, pp. 7–23. DOI: 10.1017/S1537592710004068 (cit. on p. 6).
Malmborg, Frans af (2023). "Narrative dynamics in European Commission AI policy—Sensemaking, agency construction, and anchoring". In: *Review of Policy Research* 40 (), pp. 757–780. DOI: 10.1111/ropr.12529 (cit. on p. 9).
Manheim, Karl M. and Lyric Kaplan (2019). "Artificial Intelligence: Risks to Privacy and Democracy". In: *Yale Journal of Law & Technology* 21, pp. 106–188 (cit. on p. 9).
Meaker, Morgan (2023). "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy". In: *Wired* (cit. on p. 10).
Nemitz, Paul (2018). "Constitutional democracy and technology in the age of artificial intelligence". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133. Publisher: Royal Society, p. 20180089. DOI: 10.1098/rsta.2018.0089 (cit. on p. 10).
Perrigo, Billy (2023). "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer". In: *TIME* (cit. on p. 12).
Roberts, Huw, Josh Cowls, Emmie Hine, Francesca Mazzi, et al. (2021). "Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US". In:

Science and Engineering Ethics 27.6. DOI: 10.1007/s11948-021-00340-7 (cit. on p. 11).

Roberts, Huw, Josh Cowls, Emmie Hine, Jessica Morley, et al. (2021). *Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes*. SSRN Scholarly Paper ID 3811034. Rochester, NY: Social Science Research Network (cit. on p. 11).

Robles, Pedro and Daniel J. Mallinson (2023). "Catching up with AI: Pushing toward a cohesive governance framework". In: *Politics & Policy* 51.3, pp. 355–372. DOI: 10.1111/polp.12529 (cit. on p. 11).

Roche, Cathy, P. J. Wall, and Dave Lewis (2023). "Ethics and diversity in artificial intelligence policies, strategies and initiatives". In: *AI and Ethics* 3.4, pp. 1095–1115. DOI: 10.1007/s43681-022-00218-9 (cit. on pp. 8, 12).

Rosenau, James N. (1992). "Governance, order, and change in world politics". In: *Governance Without Government*. Ed. by James N. Rosenau and Ernst-Otto Czempiel. ZSCC: 0000240. Cambridge: Cambridge University Press, pp. 1–29. DOI: 10.1017/CBO9780511521775.003 (cit. on p. 6).

Rotolo, Daniele, Diana Hicks, and Ben R. Martin (2015). "What is an emerging technology?" In: *Research Policy* 44.10, pp. 1827–1843. DOI: 10.1016/j.respol.2015.06.006 (cit. on pp. 4, 5).

Schiff, Daniel (2022). "Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy". PhD thesis. DOI: 10.17605/OSF.IO/KW8XD (cit. on p. 8).

Schiff, Daniel et al. (2021). "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection". In: *IEEE Transactions on Technology and Society* 2.1, pp. 31–42. DOI: 10.1109/TTS.2021.3052127 (cit. on p. 8).

Stix, Charlotte and Matthijs M. Maas (2021). "Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy". In: *AI and Ethics*. DOI: 10.1007/s43681-020-00037-w (cit. on p. 5).

Stone, Diane (2019). "Making Global Policy". In: *Elements in Public Policy*. DOI: 10.1017/9781108661690 (cit. on p. 7).

Tallberg, Jonas, Magnus Lundgren, and Johannes Geith (2023). *AI Regulation in the European Union: Examining Non-State Actor Preferences*. arXiv:2305.11523 [econ, q-fin]. DOI: 10.48550/arXiv.2305.11523 (cit. on pp. 8, 9, 11).

Tan, Rebecca and Regine Cabato (2023). "Behind the AI boom, an army of overseas workers in 'digital sweatshops'". In: *Washington Post*. Section: Asia (cit. on p. 12).

Ulnicane, Inga et al. (2021). "Framing governance for a contested emerging technology:insights from AI policy". In: *Policy and Society* 40.2, pp. 158–177. DOI: 10.1080/14494035.2020.1855800 (cit. on p. 9).

— (2022). "Governance of Artificial Intelligence". In: *The Global Politics of Artificial Intelligence*. 1st ed. Boca Raton: Chapman and Hall/CRC, pp. 29–56. DOI: 10.1201/9780429446726-2 (cit. on p. 8).

Veale, Michael, Kira Matus, and Robert Gorwa (2023). *AI and Global Governance: Modalities, Rationales, Tensions*. DOI: 10.31235/osf.io/ubxgk (cit. on pp. 7, 11).

Weiss, Thomas G. (2000). "Governance, Good Governance and Global Governance: Conceptual and Actual Challenges". In: *Third World Quarterly* 21.5. ZSCC: 0001138 Publisher: [Taylor & Francis, Ltd., Third World Quarterly], pp. 795–814 (cit. on p. 6).

Worthington, Richard, Mikko Rask, and Lammi Minna, eds. (2013). *Citizen Participation in Global Environmental Governance*. oth ed. Routledge. DOI: 10.4324/9781315870458 (cit. on p. 8).

Zarkadakēs, Giōrgos (2020). *Cyber republic: reinventing democracy in the age of intelligent machines*. Cambridge, Massachusetts: The MIT Press (cit. on p. 10).

Chapter 1

MAPPING GLOBAL AI GOVERNANCE: A NASCENT REGIME IN A FRAGMENTED LANDSCAPE

Abstract: The rapid advances in the development and rollout of artificial intelligence (AI) technologies over the past years have triggered a frenzy of regulatory initiatives at various levels of government and the private sector. This article describes and evaluates the emerging global AI governance architecture and traces the contours of a nascent regime in a fragmented landscape. To do so, it organises actors and initiatives in a two-by-two matrix, distinguishing between the nature of the driving actor(s) and whether or not their actions take place within the existing governance architecture. Based on this, it provides an overview of key actors and initiatives, highlighting their trajectories and connections. The analysis shows international organisations' high levels of agency in addressing AI policy and a tendency to address new challenges within existing frameworks. Lastly, it is argued that we are witnessing the first signs of consolidation in this fragmented landscape. The nascent AI regime that emerges is polycentric and fragmented but gravitates around the Organisation for Economic Co-Operation and Development (OECD), which holds considerable epistemic authority and norm-setting power.

Publication information

Title	Mapping global AI governance: a nascent regime in a fragmented landscape
Journal	AI and Ethics
Publication date	17 August 2021 (online)
Citation	Schmitt, L. Mapping global AI governance: a nascent regime in a fragmented landscape. AI Ethics 2, 303-314 (2022)
URL	`https://link.springer.com/article/10.1007/s43681-021-00083-y`
DOI	`https://doi.org/10.1007/s43681-021-00083-y`
Keywords	Global governance; Artificial intelligence; Global AI governance; AI policy; AI ethics

Note: The following chapter is equivalent to the published article. Spelling has been adapted to British English to align the text with the overall dissertation.

1.1 Introduction

The rise of artificial intelligence (AI) technology and its transformative impact across a wide range of issues pose new challenges to policymakers and other stakeholders around the globe. Whether one looks at the near, medium, or long term, there arise a myriad of legal and ethical challenges and even existential risks that societies need to address. These risks are exacerbated by a lack of effective global governance mechanisms to provide, at minimum, guardrails steering AI in beneficial directions (Cihon 2019; Gill 2020).

Taken together, the unprecedented advances in AI development and deployment over the past years led to a challenging and rapidly evolving research agenda of issues that touches upon various aspects of the digital ecosystem. Researchers look at questions pertaining to ethics, regulation, and governance, with a strong normative push to ensure that the potential malign consequences of AI are controlled and the benefits fairly distributed (MacIntyre, Medsker, and Moriarty 2021; Bostrom and Yudkowsky 2014; Veale, Van Kleek, and Binns 2018; Cave et al. 2019; Turchin and Denkenberger 2020). Important conceptual and technical work sheds light on foundational issues like algorithmic transparency, explainability, or safety, as well as on related aspects such as data governance or privacy (Thelisson, Padh, and Celis 2017; Lepri et al. 2018; Miller 2019; Brundage et al. 2020). Similarly, the ethical aspects and their implementation in code, as well as organisational governance, have received many valuable contributions (Thelisson, Padh, and Celis 2017; Gasser and Almeida 2017; Thelisson, Morin, and Rochel 2019; *Perspectives on Issues in AI Governance* 2019; Kurshan, Shen, and Chen 2020).

However, as the AI ethics research community develops frameworks and technical governance models to ensure that AI is designed and employed ethically, it must not forget about the global dimension of AI policy[1] (ÓhÉigeartaigh et al. 2020; Rotenberg 2019). Values-based approaches, ethics-by-design, and other principled suggestions must also be translated into a functional system of rules, binding agreements, and international governance mechanisms that go beyond voluntary self-commitments or hollow AI strategies.

Partially, this work will take place at the national or even sub-national level. However, to a large extend, AI policy will be shaped internationally. Cutting-edge AI research is already a global enterprise dominated by large transnational technology firms. Moreover, the cross-border nature of the digital ecosystem renders purely national regulatory regimes inefficient and costly. Hence, big parts of the discussion around ethical AI and AI governance, just as this article, focus on the international level.

Furthermore, AI development does not happen in a vacuum, and as it gains salience on the political agenda, this also results in geopolitical, strategic considerations taking

[1] AI policy describes soft or hard governance measures which may take a range of forms such as principles, codes of conduct, standards, regulatory or legislative approaches (Stix and Maas 2021).

over the way that governments around the world position themselves towards potential regulatory or legislative measures (Gill 2019; Imbrie et al. 2020). Global governance scholarship provides a useful lens to understand and explain these developments.[2]

Therefore, this article sets out to describe and evaluate the rapidly evolving emergent global AI governance landscape. To do so, it organises actors and initiatives in a two-by-two matrix, distinguishing between the nature of the driving actor(s), and whether these take place within the existing architecture or instead create novel instruments. As the overview and subsequent analysis shows, multiple initiatives compete for influence in a fragmented landscape. Many of these are state-led, but international organisations have demonstrated a surprising level of agency in addressing AI policy. And even though AI is a novel technology going beyond the scope of established regulatory or legal governance mechanisms, there is a tendency to address these new challenges within existing frameworks. The final section discusses the findings and argues that we begin to see signs of consolidation of a nascent AI regime that is polycentric and fragmented but gravitates around the OECD, which holds considerable epistemic authority and norm-setting power.

1.2 Literature review

The rich and fast-growing scholarship on AI ethics often touches upon political questions and issues of international relations (Gasser and Almeida 2017). Nevertheless, literature specifically addressing global AI governance or international AI policy is rare to find. This article aims to contribute to recent work on the important intersection of AI ethics and governance (ÓhÉigeartaigh et al. 2020; Schiff et al. 2020; Shi 2020).

In outlining a research agenda for AI governance, Dafoe (2018) laments the neglect of political science in understanding problems of AI governance. He emphasises the role of the discipline in shaping AI politics and devising visions for ideal governance. This notion is echoed by Parson et al. (2019), who put a focus on the social processes by which AI technologies are developed and applied. The most related to the present article is an instructive overview by Butcher and Beridze (2019), in which the authors provide an overview of current AI governance activities. They cover many notable examples across the private sector, public sector, research and multi-stakeholder organisations, and the UN. AI development is fast-paced, however, and so is the political environment in which AI governance is shaped. Hence, the present article's updated and expanded mapping presents a vital contribution to the literature.

[2] A key distinguishing feature of global governance is its incorporation of a plurality of actors, both private and public, state and non-state, who engage at multiple levels and through their interactions define the global order. In this view, the term 'global' is a clear delineation from traditional international relations literature, which are thought of as too state-centric. Likewise, 'governance' stands in contrast to government, speaking to a wider range of actors and a more fluid understanding of power hierarchies.

Other recent work has covered the topic from various angles, such as the role of international standards (Cihon, Maas, and Kemp 2020a), national AI strategies (Campbell 2019) or ethics guidelines (Schiff et al. 2020; Larsson 2020). Valuable contributions have also been made theorising about the governance design of international agreements (Stix and Maas 2021; Cihon, Maas, and Kemp 2020b).

The latter contributions have shifted focus away from observing existing governance architectures to analysing and theorising how they ought to be designed. This strand of research is more prescriptive and directly feeds into important policymaking considerations. One of the earlier contributions in this space investigates the role and competencies of different institutions in managing or regulating AI, even proposing a regulatory regime for AI (Scherer 2015). However, this and similar work focus on the national rather than the international level. Researchers have also proposed specific initiatives for the US to foster international cooperation on AI (Andrew Imbrie et al. 2020; Rasser et al. 2019) or even set up new international bodies (Prakash 2019). Looking explicitly at global governance institutions, scholars have recently drawn up a leadership role for the G20 in defining global public policy on AI (Pomares and Abdala 2020; Jelinek, Wallach, and Kerimi 2020; Abdala, Ortega, and Pomares 2020).

These treatises of ideal AI governance are certainly important. Yet, there is still a great gap in better describing and theorising the current state of AI governance. Researchers need to study the legal and governance responses to AI, "both within traditional legal and regulatory settings, and in new institutional mechanisms and settings" (Parson et al. 2019). Reaching a deeper understanding of the political dynamics at the global level will help answer questions of how to move from the current to the ideal.

In light of this, the contribution of this article is three-fold: (1) the two-by-two matrix provides a useful analytical tool for organising and evaluating the current governance landscape, serving as a reference point; (2) the empirical work on important governance actors and initiatives, together with the analysis of their trajectories and connections, helps answer important questions about the dynamic evolution of AI governance; (3) the discussion of these findings brings to light important features and patterns of the nascent AI regime. All this, ultimately, allows to extrapolate and anticipate future directions of travel and ask questions to guide further research.

1.3 Mapping the current global AI governance landscape

The current global AI governance landscape displays a multitude of governance initiatives by various actors, some dealing with the regulation of very specific AI applications and others with more general, abstract principles of AI ethics and policy. By now, many countries have brought forward their own AI strategies, often with direct reference to the international level and questions of global AI governance. While these alone are

valuable objects of study, this article focuses on those actors and initiatives that are by nature transnational or multilateral, i.e., that involve stakeholders from more than two countries. At this stage, almost none of them entail binding legislation, but rather political declarations, ethical principles, or partnerships.

There are many ways to organise these actors and initiatives, structuring them by their regional or topical scope, by the actors' nature (e.g., governmental, business, civil society) or by the kind of instrument involved (e.g., international treaty or organisation, alliances or partnerships, political declarations). This article employs a two-by-two matrix (table 1.1) that distinguishes the following characteristics: (a) between action that is embedded in the existing governance architecture vs action that establishes new instruments; and (b) between state-led initiatives and non-state-led initiatives.[3] Note that the latter dimension considers the origin or agency of action, not necessarily the organisational nature through which it is ultimately carried out.[4]

	State-led	**Non-state-led**
Embedded in existing architecture	- G7 - G20 - CCW Group of Governmental Experts on emerging technologies in the area of LAWS (GGE) - Council of Europe (CoE)	- United Nations (UN) - European Commission - Organisation for Economic Co-operation and Development (OECD) - IEEE - ISO/IEC
Establishing new instruments	- Global Partnership on AI (GPAI) - AI Partnership for Defense	- Partnership on AI (PAI)

Table 1.1: *An overview of the most important multilateral governance initiatives and actors in the AI domain.*

This overview is by no means comprehensive. There are dozens of other actors

[3] The non-state group includes supranational actors such as international organisations. Hence, the state- vs non-state distinction is not one of public/governmental vs private. Rather, it is intended to show how the nation-state as the traditional actor in international relations compares to other possible actors as understood by global governance theory.

[4] Clearly, these categories are somewhat arbitrary and even the assignment into the categories is open for interpretation. To illustrate these choices, consider the following: The GGE is within the UN framework, yet it is mainly driven by national governments. Similarly, the G7 and G20 may be considered as informal international organisations, yet they have no agency of their own and merely express the preferences of the participating nations. Therefore, these are put in the left-hand category. In contrast, international organisations with independent secretariats or executive bodies, such as the UN Secretariat-General, the European Commission, and the OECD, are found on the right-hand side. While they may ultimately respond to the governments of their member states, they have demonstrated a substantial degree of autonomy and agency in driving their respective AI policies. Note that the Council of Europe (CoE) does not sit on the right-hand side, despite being an international organisation. Unlike the previous organisations, the CoE's work on AI was directly mandated by its member states, hence state-led.

and initiatives engaging in AI governance. However, the present sample confidently includes the most important ones to date, disregarding others (see section 1.3.5) for the sake of feasibility and analytical precision.[5]

The following section briefly summarises and contextualises the different actors and initiatives, highlighting their trajectories and connections. This exercise is mostly agnostic to the content of what these global governance initiatives and arrangements actually entail. Focusing the analysis on actors and instruments was a deliberate choice to avoid confusion between structure and content.

The fragmented landscape that emerges from this exercise is congruent with other authors' characterisations of the nascent global AI governance architecture as an "unorganised" and "immature field" (Butcher and Beridze 2019). Alongside the different actors, we find epistemic communities (Haas 1992) that are well-connected and often overlap. Governance actors differ in their agenda-setting and norm-setting powers (Schiff et al. 2020). The analysis also shows the rapid progress and first signs of consolidation and convergence. Furthermore, the observed dynamics shed some light on the type of entities involved in the early design of AI governance – which is marked more by the utilisation of existing governance instruments than by institutional innovation. In addition, it demonstrates a surprisingly high level of agency by international organisations. These findings are discussed in more detail in the last section.

1.3.1 State-led initiatives embedded in the existing architecture

The subset of "state-led initiatives embedded in the existing architecture" includes four cases presented below. The timeline of developments gives testament to the often slow and laborious processes of international diplomacy: Governments started to treat AI-related policy challenges seriously within international fora beginning in 2016. Since then, the topic has risen in priority, as shown by its ascent from the ministerial to the leaders' level over time.

G7

The G7 has been a popular forum for leaders of some of the largest democracies to discuss AI issues. Initially, discussions were held at the level of ministerial meetings but were later brought up to the leaders' level. The G7 ICT Ministerial 2016 in Japan and 2017 in Italy resulted in a statement outlining a vision of human-centric AI for innovation and economic growth. Then, in March 2018, the G7 innovation ministers agreed on a "Statement on Artificial Intelligence." Building on this statement, the Canadian

[5] The author is aware of the potential for Western and Eurocentric geographic and linguistic selection biases, which may result in the neglection of relevant observations. Feedback and suggestions to correct for possible blind spots are highly welcome.

G7 presidency hosted the "G7 Multistakeholder Conference on Artificial Intelligence" in December 2018, convening over 200 AI experts from the G7 countries and beyond.

The most notable action at the leaders' level was when in June 2018, the G7 committed to the Charlevoix Common Vision for the Future of Artificial Intelligence. It includes 12 commitments to promote human-centric AI fostering economic growth, societal trust, and equality and inclusion.

Since then, the most notable development within the G7 framework has been the inception of the Global Partnership on Artificial Intelligence (GPAI). Because it quickly expanded beyond the G7 both in membership scope and organisational structure, it is listed amongst the "state-led initiatives creating new instruments" and will be discussed in more detail in the next section.

G20

Trailing the course of the G7, the G20 got active on AI policy a little later. In June 2019, under Japanese leadership, members agreed on a ministerial statement that focused on human-centred AI. Even more noteworthy, they endorsed the OECD's set of principles on trustworthy AI (see next section). This can be seen as a major achievement of the OECD, who this way expanded its reach to some of the major players outside its membership base (esp. China and Russia).

The commitment to advance the G20 AI principles was confirmed in 2020 by its digital ministers, under the Saudi Arabian presidency. At the meeting, countries also collected examples of national strategies and policy initiatives aimed at trustworthy AI. Promisingly, even China seems to have fully subscribed to the G20 AI principles, as a speech by Xi Jinping delivered at the G20 leaders' summit in November 2020 indicates (Xinhua 2020).

These signs of activism at the G20 have led some observers to call for its primacy in global AI governance (Pomares and Abdala 2020) and the establishment of a "G20 coordinating committee for the governance of artificial intelligence" (Jelinek, Wallach, and Kerimi 2020). However, no such progress has materialised to date.

CCW GGE

Targeting only a specific application of AI, namely lethal autonomous weapons systems (LAWS), the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE) meets since 2017 within the framework of the United Nations Convention on Certain Conventional Weapons (CCW). With representatives from on average 90 states (Reaching Critical Will 2020), the GGE can be considered the broadest international forum for talks on issues directly related to AI applications. The meetings usually take place once per year for a 1-week-session and have brought to the table the most relevant states.

A first breakthrough was achieved in 2018 when members identified ten guiding principles related to important aspects such as human responsibility and international humanitarian law. Yet, discussions have proven to be slow and difficult due to the lack of consensus on agenda items. A coalition of civil society organisations and countries have called for a legally binding instrument banning LAWS. However, given the continued resistance of major players such as the US, France, and Russia, a breakthrough towards an effective governance framework seems unlikely at this stage (Barbé and Badell 2020). Also, the unique characteristics of AI-enabled weapon systems have some observers wondering whether traditional approaches to arms control and disarmament are suitable at all (Gill 2019).

Council of Europe

The Council of Europe (CoE) has made splashes with its foray into AI governance. In September 2019, the CoE's executive body – consisting of the member states' foreign affairs ministers – established the Ad-Hoc Committee on AI (CAHAI). It was tasked with "examining, through broad multistakeholder consultations, the feasibility and potential elements of a legal framework for the development, design, and application of AI" (Council of Europe 2020).

In addition to bringing together member and observer states' views as well as input from civil society, academia, and the private sector, the CAHAI is also cooperating closely with other international institutions, such as UNESCO, the OECD and the European Commission.

CAHAI released a comprehensive collection of government contributions in its interim report (Ben-Israel et al. 2020) – which curiously was funded by Japan, itself not a member of the CoE (but an observer since 1996). In the report, the organisation reaffirms its "clear role to address the issue of the development and uses of AI" and proposes to work towards a horizontal legal instrument whose principles could serve as a basis for more specialised texts. Whether such an instrument will eventually materialise remains to be seen – the roadmap foresees a final report and a decision by member states by the end of 2021.

In any case, the CAHAI already developed into an important forum developing knowledge and stimulating exchange between almost 50 states. Since the membership includes a diverse set of actors with at times opposing interests regarding AI development (e.g., Russia is a member of the CoE as well), agreement on actual binding outcomes seems unlikely – though this would no doubt be a substantial step towards global AI governance.

1.3.2 State-led initiatives creating new instruments

The subset of "state-led initiatives creating new instruments" is substantially smaller than the previous one. It consists of only two initiatives, both of which are predominantly driven by developed democracies.

GPAI

The first one is the Global Partnership for AI (GPAI), as mentioned above. The GPAI was originally introduced in 2017 by Canada and France under a different name (International Panel on Artificial Intelligence). The initial response was timid, and the proposal long-faced strong reluctance from the Trump administration over concerns that moves towards any sort of regulation might hamper innovation in AI. Finally, in May 2019, the US changed course, now considering the GPAI as a useful tool in restricting China's influence on the emerging global AI governance system (OBrien 2020).

GPAI was officially launched in June 2020 with a total of 15 founding members.[6] By December 2020, Brazil, the Netherlands, Poland and Spain had joined the partnership, underscoring the partnership's appeal and potential for expansion. The GPAI's stated aim – grounded in human rights, inclusion, diversity, innovation, and economic growth – is to guide the responsible development and use of AI. By bringing together experts from industry, government, civil society and academia, it hopes to facilitate international collaboration and act as a global reference point for specific AI issues (Plonk 2020). Importantly, it also adheres to the OECD's Principles on AI, another sign of the OECD's successful role as a global norm-setter.

The GPAI's evolution is an interesting case: conceived within the existing architecture, it was then launched as a separate, standalone initiative with a unique membership base going beyond the G7. Thus, it demonstrates characteristics of a new, standalone instrument, while ultimately ending up hosted by one of the existing international organisations (the OECD).

The GPAI is arguably the most advanced global AI governance instrument to date, with a permanent secretariat and a relatively broad membership base. While it is so far missing major players such as China and Russia, an incoming Biden administration could potentially make the US more inclined to finding consensus and thus broaden the alliance further. In any case, the OECD's role as a host of the GPAI will be useful in avoiding policy incoherence and fragmentation, given that the OECD is also closely aligned with the G20 (see above).

[6] Australia, Canada, France, Germany, India, Italy, Japan, Mexico, New Zealand, the Republic of Korea, Singapore, Slovenia, the United Kingdom, the United States and the European Union.

AI Partnership for Defense

Another recent state-led initiative is the US-driven AI Partnership for Defense. Six NATO members as well as other US allies such as Israel, Japan, and Sweden, followed an invitation by the Pentagon's Joint Artificial Intelligence Center to a virtual conference in September 2020. Discussions ranged from policy issues such as ethical principles to military use cases of AI and scope for technical cooperation (Freedberg Jr 2020). It remains to be seen how this lose group continues under a Biden administration, but in any case, the large response by partners signals the growing interest for cooperation in security and defence matters related to AI. This is also backed by statements from high-ranking NATO officials which have supported increased transatlantic cooperation on this matter in the past (NATO 2020).

1.3.3 Non-state-led initiatives embedded in the existing architecture

Global governance literature brought to the international relations scholarship a stronger focus on non-state actors, ranging from the important role of international organisations and businesses to civil society actors and non-governmental organisations (Zürn 2018; Jordana and Triviño-Salazar 2019). In this vein, this mapping of the emergent global AI governance landscape would not be complete without looking at non-state-led initiatives.

This section fields several international organisations, namely the UN, EU, and the OECD. All three organisations have actively sought to take over leadership roles in the previously unoccupied space of AI governance. This section also includes some international standard-setting organisations, who are competing over the formulation of a variety of standards – from narrowly technical to more general and political – in the AI domain.

UN

The most obvious candidate to look at when describing any global governance system is the United Nations (UN). Secretary-General António Guterres has emphasised the impact of emerging technologies, including AI. In 2018, he established a High-Level Panel on Digital Cooperation, a multi-year, multi-stakeholder, global effort to address a range of issues related to the Internet, artificial intelligence, and other digital technologies. Its results were presented as a "Roadmap for Digital Cooperation" in June 2020 and included a recommendation on global AI cooperation for AI that is "trustworthy, human-rights based, safe and sustainable and promotes peace" (United Nations 2020). In the roadmap, Guterres states his intention to establish a multi-stakeholder advisory body on global AI cooperation, comprising member states, relevant UN entities, interested companies, academic institutions, and civil society groups. The body should

"serve as a diverse forum to share and promote best practices, as well as exchange views on artificial intelligence standardization and compliance efforts." Besides, he committed to appointing an Envoy on Technology by 2021. It remains to be seen whether these intentions will be followed up on in 2021, and whether their outcomes will match expectations.

Other parts of the UN system are also becoming engaged in AI governance: already in 2015, the UN Interregional Crime and Justice Research Institute (UNICRI) launched a programme on AI and robotics. That same year, AI governance was discussed for the first time during the 70th UN General Assembly (Butcher and Beridze 2019). The UN Institute for Disarmament Research (UNIDIR) supports the work of the GGE on LAWS. UNIDIR and the UN University Centre for Policy Research (UNU-CPR) have also set up research projects to explore AI-related policy challenges.

More and more UN agencies are looking at AI, both as a disruptive technology to their respective policy domain and as a tool to achieve the Sustainable Development Goals (ITU 2020). For instance, since 2017, the International Telecommunication Union (ITU) co-organises an annual AI for Good Global Summit. Then there is UN Global Pulse, the Secretary-General's initiative on AI for humanitarian aid and development, which also looks at AI governance. Amongst its work, it convenes an Expert Group on Governance of Data and AI, bringing together international leaders from the public and private sector, civil society, and the legal community (Pizzi and Romanoff 2020). All these efforts go hand in hand with the wider AI for Social Good (AI4SG) movement, aimed at establishing interdisciplinary partnerships centred around AI applications towards SDGs (Tomašev et al. 2020).

These various efforts establish the UN as a global convening platform for stakeholders interested in exploring how AI can contribute to achieving the SDGs and solve global problems. This gives the UN considerable epistemic authority, though somewhat undermined by the multitude of initiatives and work streams causing inconsistency and complexity. It also allows it to support consensus-building between states to promote common goals, thus making AI governance more effective (Butcher and Beridze 2019).

European Commission

The European Commission took action on AI policy even before most EU member states did. With the release of its AI strategy in April 2018, it established the High-Level Expert Group on Artificial Intelligence, whose 52 members have been at the fore of global debate around AI regulation and governance questions. The Commission furthermore chartered new ground with the release of its widely-noted 'Ethics Guidelines for Trustworthy AI' in April 2019.

Since then, the Commission published a White Paper on AI in February 2020 as a preparatory step for a forthcoming legislative proposal on AI. During a month-long

open consultation process, Brussels has probed reactions to its proposed approaches for regulating and governing the technology's development and application. The gathered feedback should be transformed into a legislative proposal that is expected in the first half of 2021 and setting the EU on track to become the first major jurisdiction worldwide with a binding legal framework for AI.

The Commission's overall goal is to chart a so-called "European third way" for AI development, which policy-makers frame as "human-centric", "ethical", and "trustworthy." If (or rather, when) translated into hard law, this will undoubtedly have repercussions well beyond the EU's direct jurisdiction, as has been demonstrated by the global reach of the EU's privacy directive GDPR.[7]

In addition to these strategic, regulatory and legislative approaches, the European Commission acts in close coordination with its member states, mainly through the Coordinated Plan on AI. It also liaises with the wider AI community and especially with industry, bringing together over 4000 representatives via its European AI Alliance – a multi-stakeholder forum launched in June 2018. On the global level, it is engaging with most other actors listed in this overview. Moreover, it is a founding member of GPAI, underscoring its active role in the international arena.

OECD

Another well-known international organisation that has sought ownership of AI-related governance issues is the Organisation for Economic Co-operation and Development (OECD). Back in 2016, the OECD's Committee on Digital Economy Policy began discussing the need for AI principles and established an expert group in May 2018. The resulting "OECD Principles on AI" were adopted in May 2019 as the "first set of intergovernmental policy guidelines on AI" and included commitments to trustworthy, human-centred AI (OECD 2021). Beyond the OECD members, Argentina, Brazil, Costa Rica, Malta, Peru, Romania, and Ukraine have also signed up to the AI principles, signalling broad international appeal.

In addition to this work, the OECD built up a considerable public-access knowledge base called the OECD.AI Policy Observatory, launched in February 2020, to help policymakers implement the AI principles and further inform the global discourse on AI governance. Also, the OECD Network of Experts on AI (ONE AI), a multi-disciplinary and multi-stakeholder group, was set up to provides AI-specific policy advice and foster international cooperation.

These efforts paid off: when France and Canada used their G7 presidencies and together with 13 other founding members launched the GPAI as discussed in the previous section, they decided to host its secretariat at the OECD. This hybrid structure has the potential to foster synergies between the OECD-led work on global AI policy and the

[7] For more on how this so-called 'Brussels Effect' (Bradford 2012) may play out with regards to the EU's role in global AI governance, see (Schmitt 2020).

GPAI's more technical discourse (Plonk 2020). Furthermore, as mentioned above, the OECD's principles were endorsed by the G20, which includes China and Russia, thus giving it an even broader international reach. The principles also serve as the basis for the work of the GPAI, thus anchoring the alliance firmly within the OECD's sphere of influence – both organisationally (hosting of the secretariat) as well as normatively.

Granted, the OECD does not have any regulatory or legislative power, including on AI policy. In any case, binding international treaties that regulate the development and use of AI horizontally seem far-fetched at this point. What remains is soft power – the ability to influence global AI governance through epistemic authority, convening power, and norm- and agenda-setting. In this realm, the OECD has demonstrated considerable strength.

Standards organisations

The subset of "non-state-led initiatives embedded in the existing architecture" also includes international standard-setting organisations, whose membership base is usually dominated by industry and business associations.[8] In the following, the article describes the ongoing work on AI standards at the two leading international standards bodies. Their work tends to be rather technical. However, standards – and especially international standards – undoubtedly affect the development and roll-out of AI technology and, by extension, the corresponding regulatory and governance domains.[9] Furthermore, the epistemic authority of these standard-setting bodies informs and influences policymaking by other actors directly and indirectly.

The International Organization for Standardization (ISO) together with the International Electrotechnical Commission (IEC) have a dedicated AI sub-committee, ISO/IEC JTC 1/SC 42, since 2017. While most of its work is dedicated to technical aspects, it explicitly frames its work on AI standardisation as a new "holistic ecosystem" approach that also considers ethical and societal concerns (Diab 2020). Its members actively engage with other relevant global stakeholders such as the OECD, the European Commission, and the Partnership on AI (ibid.).

Furthermore, important efforts are being made at the IEEE Standards Association at least since 2016. Its "IEEE Global Initiative on Ethics of A/IS" aims at, inter alia, "global consensus building to inspire the Ethically Aligned Design of autonomous and intelligent technologies" (IEEE 2021). Out of this initiative have sprung several relevant work strands and publications.[10] Grouped under the IEEE P7000 standards family, 14

[8] In some cases, such as the with the ISO, the international umbrella organisation is composed of national member bodies. These, in turn, tend to be largely industry-based.

[9] For a detailed treatment of the work of international standards bodies and their impact on AI governance, see Cihon (2019).

[10] This includes the release and updates of a landmark report called "Ethically Aligned Design (EAD): A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems", first published

AI-related standards are currently being developed in Working Groups.

Lastly, the IEEE initiative has also contributed to the establishment of the Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS) in July 2018. Gathering more than 70 organisations, it is devised as a global forum for exchange and collaboration in the ethical development and use of AI-related standards.

As these organisations continue to develop standards, they will undoubtedly shape the development and use of AI. With their institutional capacity to achieve expert consensus and then promulgate standards internationally, which then become enforced either de facto or de jure, they exercise certain norm-setting powers. Their role as a global reference point, and their intensive exchanges with other global governance actors, also gives them considerable epistemic authority as well as convening and agenda-setting power.

1.3.4 Non-state-led initiatives creating new instruments

PAI

Unlike its state-led namesake GPAI, the Partnership on AI (PAI) was born out of an alliance of non-state actors, that is big American tech companies at the forefront of AI development. It was established in late 2016 by a group of AI researchers representing Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft. One year later, this business-centred setup was expanded to include six not-for-profit board members, thus turning the PAI into a multi-stakeholder organisation, which today convenes more than 50 member entities.

PAI is actively supporting research on many pressing issues related to AI ethics and governance. Besides acting as a convener and knowledge incubator, it also facilitates educational projects as well as practical tools such as the recently launched AI Incident Database (AIID). Since November 2020, the AIID documents failures of AI systems around the world. The idea of this publicly available repository is to disseminate knowledge and improve the safety of AI systems deployed in the real world. The AIID is inspired by incident databases in the aviation and computer security industries. The usefulness of such databases is undisputed; it allows developers to learn from their peers' mistakes and opens up research avenues for external observers who can thus gauge the AI world both for episodic and systemic risks. The AIID is still in its infancy and it is too early to tell whether the wider AI community accepts it as a tool of reference. However, its early-mover advantage and broad membership base enable the PAI to establish itself as a cornerstone in the emerging AI governance landscape. Since AIID is developed as an open-source project collectively governed by the PAI, it reminds us of the origins of Internet governance (Mueller, Mathiason, and Klein 2007). Whether the

in 2016. In the document, AI developers are encouraged to prioritise ethical considerations in the creation of autonomous and intelligent technologies.

AI community will continue in this collaborative path remains to be seen. We simultaneously observe signs that as the AI industry matures, it also increasingly moves to proprietary models and favours commercial over common interest (Hao 2020; Quach 2019).

1.3.5 Other actors: NGOs, research institutes, and global movements

Besides the above-mentioned actors and initiatives, there are dozens – if not hundreds – that also affect global AI governance in one way or another. These include non-governmental organisations, research institutes, public sector entities (e.g., cities and regional governments) or global movements (e.g., Campaign to Stop Killer Robots). Taken together, these provide a considerable epistemic source and their engagement in agenda-setting should not be understated. Their regular interactions with the actors discussed in previous sections can indirectly influence outcomes at the global level.

Nevertheless, their individual impact is comparably low and hence outside the scope of this study. Causes may be either that their approach to AI governance is too specific (i.e., focusing on only one specific aspect or sector of AI) or too tangential (i.e., initiatives addressing the wider digital ecosystem and only mentioning AI in passing), or simply that they lack the political clout to make their voices heard. This last point especially speaks to the important debate about inclusivity and participation in AI governance (Ashok 2017; Budish, Gasser, and Ashar 2018).

1.4 Discussion

The preceding overview of the global AI governance landscape allows for several relevant observations, which are discussed in the following section.

First, there is a clear tendency to accommodate governance initiatives within the existing architecture, both by state and non-state actors. This could have several potential explanations. States and other global governance actors might be wary of foundational innovation and starting from scratch. Instead, they prefer to build on existing, proven governance arrangements. Alternatively, more attempts might have been made with new instruments and these might simply have been less fruitful and thus did not feature in this overview. In any case, the case of the GPAI suggests a gravitational pull towards established governance mechanisms.

Second, there is a fairly equitable distribution of labour between national governments (state-led) and international organisations (non-state-led). The community of international organisations moved early to occupy an open policy space, thus carving out a considerable competence vis-à-vis its member states. These, in return, offloaded some of the AI policy work to international organisations (CoE, OECD via GPAI).

This would suggest that states accept their role as useful fora for international cooperation and the steering of AI development into globally beneficial directions. However, global coordination in this realm has so far not touched upon legally binding treaties. It may well be that governments decided to transfer some authority to IOs only as long as they deal with rather abstract principles or soft governance, but would withdraw or stall as soon as work proceeds towards more regulatory, hard governance. Whether the CoE will produce any meaningful conclusions by the end of the year may be a good indication of the potential for such binding international rules.

Thirdly, international standards organisations play a role in the development of AI governance, as is the case for most emerging technologies. More worrying is the shift towards geopolitics: in the last years, the development of international AI standards has increasingly received attention from key governments such as China, the EU, and the US. Their renewed interest and subsequent strategic engagement risks contention and the encroachment of geopolitical considerations into domains that ought to be technical (Seaman 2020; Blancato 2019). This may not only affect the quality of standards but also obstruct debates around AI ethics. As standards cannot be completely detached from the policy world, scholars of global AI governance need to have a sound understanding of the proceedings in the international standard-setting arena. Future research should explore the interactions and means by which governments aim to steer the development of standards to further their own perceived interests.

Lastly, sub-state actors from the public sector are practically not present in the discussions around global AI governance. This is in stark contrast to other policy domains such as global climate change governance, where city networks play an important role. It is also a bit surprising, given that cities are one of the focal points of AI rollout and several cities have subsequently taken notable actions with regards to AI policy. However, to date, these actions are isolated and do not engage at the supranational or global level.

In light of the fuzzy nature of AI, it is barely surprising that the current landscape is somewhat fragmented. Promising moves towards some degree of centralisation and coordination are found in the prominent role of the OECD. With its epistemic authority and its norm- and agenda-setting power, it managed to act as a reference point for the G7 and G20. Through its close collaboration with other multilateral actors such as the European Commission, the UN, and the CoE, and by using the GPAI as a dedicated tool for advancing global AI governance, it may continue to play a leading role.

With all this in mind, this article argues that we are witnessing the first signs of consolidation in this fragmented landscape. The nascent AI regime that emerges is polycentric and fragmented but gravitates around the OECD, which holds considerable epistemic authority and norm-setting power. It is polycentric because it features different epistemic communities and multiple centres of decision-making, each operating with some degree of autonomy. It is fragmented because there is substantial overlap in different actors' membership and the topics addressed by these initiatives; the well-connected

epistemic communities are equally overlapping. As with other polycentric governance architectures, global AI governance will likely continue to struggle with the challenge of coordination (Carlisle and Gruby 2019). While epistemic and membership overlap may benefit consolidation or convergence, topic overlap tends to foster fragmentation and adds complexity to the regime.

This article has been mostly agnostic to the content of what these global governance initiatives and arrangements actually entail. It was a deliberate choice to focus the analysis on structure, actors and instruments, to avoid confusion between structure and content. Nevertheless, a quick look at the main developments suggests that there is convergence on a certain type of AI values and principles, as put forward by the European Commission and the OECD. These are focusing on trustworthy, human-centric AI.

Such terms are of course abstract and somewhat vague, thus leaving room for interpretation. This interpretation, contextualisation, and operationalization of AI values will without doubt experience major contestation by different actors. While China is side-lined from most of the above initiatives, its role in AI governance cannot be understated. The government has signalled willingness to engage in global governance as a responsible actor, and specifically on AI ethics has made some steps towards conciliation. Yet, it will want to interpret AI ethics in accordance with its own cultural context and promote these views globally. Hence, how China engages with the GPAI and other governance initiatives (and vice-versa) will be an interesting space to watch and leaves ample room for future research.

1.5 Conclusion

This article outlined the current state of play in global AI governance by describing the most important multilateral initiatives. It thus contributes to the growing body of literature aimed at understanding and engaging with the rapidly evolving global AI governance architecture. It organised individual actors and initiatives in a two-by-two matrix, distinguishing between the nature of the driving factor(s) and whether or not their actions take place within the existing governance architecture. Based on this, it provided an overview of key actors and initiatives, highlighting their trajectories and connections. Lastly, it has been argued that we are witnessing the first signs of consolidation in this fragmented landscape. The nascent AI regime that emerges is polycentric and fragmented but gravitates around the OECD, which holds considerable epistemic authority and norm-setting power.

The analysis has traced interlinkages and sequential developments which shed additional light on the evolving nature of this dynamical field. It also brought to light valuable insights into the emergent governance regime: most interactions are accommodated within the existing governance architecture, such as the UN system, established

IOs, and international standard-setting bodies. Supranational organisations such as the EU and the OECD have demonstrated remarkable agency in shaping global AI governance.

Whether these observations parallel developments in other global governance architectures (e.g., climate, nuclear safety, or Internet) would be an interesting avenue for future research. Also, the complementary future analysis could look in more detail at the strategies and actions of nation-states and how these engage on the global level in shaping AI governance.

Building on such descriptive empirical work, further research on global AI governance could engage more thoroughly with analytical and theoretical questions. We might ask, for instance, how this nascent global AI governance system fits into the wider global governance architecture (see Biermann et al. 2009). Or, in the absence of singular central authority in the global AI governance system, what polcycentricity theory (Ostrom 2010; Jordan et al. 2018) can tell us about the way in which actors mutually adjust and order relationships with one another.

Bibliography

Abdala, María Belén, Andrés Ortega, and Julia Pomares (2020). *The Future of Multilateralism and Global Governance*. Tech. rep. T20, p. 15 (cit. on p. 21).

Ashok, Aparna (2017). *Top takeaways from the Global Symposium on AI and Inclusion* (cit. on p. 32).

Barbé, Esther and Diego Badell (2020). "The European Union and Lethal Autonomous Weapons Systems: United in Diversity?" In: *European Union Contested*. Ed. by Elisabeth Johansson-Nogués, Martijn C. Vlaskamp, and Esther Barbé. Cham: Springer International Publishing, pp. 133–152 (cit. on p. 25).

Ben-Israel, Isaac et al. (2020). *Towards Regulation of AI Systems - Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law*. Tech. rep. DGI (2020)16. Council of Europe (cit. on p. 25).

Biermann, Frank et al. (2009). "The Fragmentation of Global Governance Architectures: A Framework for Analysis". In: *Global Environmental Politics* 9.4, pp. 14–40. DOI: 10.1162/glep.2009.9.4.14 (cit. on p. 35).

Blancato, Filippo Gualtiero (2019). *Regulate to Dominate: The Geopolitics of Standard-Setting in Digital Technologies and its Strategic Implications for the EU*. Policy Brief 8. United Nations University Institute on Comparative Regional Integration Studies (cit. on p. 33).

Bostrom, Nick and Eliezer Yudkowsky (2014). "The ethics of artificial intelligence". In: *The Cambridge Handbook of Artificial Intelligence*. Ed. by Keith Frankish and William M. Ramsey. Cambridge: Cambridge University Press, pp. 316–334. DOI: 10.1017/CBO9781139046855.020 (cit. on p. 19).

Bradford, Anu (2012). *The Brussels Effect*. SSRN Scholarly Paper ID 2770634. Rochester, NY: Social Science Research Network (cit. on p. 29).

Brundage, Miles et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims". In: *arXiv:2004.07213 [cs.CY]*. DOI: https://doi.org/10.48550/arXiv.2004.07213 (cit. on p. 19).

Budish, Ryan, Urs Gasser, and Amar Ashar (2018). *Global Governance and Inclusion* (cit. on p. 32).

Butcher, James and Irakli Beridze (2019). "What is the State of Artificial Intelligence Governance Globally?" In: *The RUSI Journal*, p. 10 (cit. on pp. 20, 23, 28).
Campbell, Thomas A (2019). "Artificial Intelligence: An Overview of State Initiatives". In: p. 45 (cit. on p. 21).
Carlisle, Keith and Rebecca L. Gruby (2019). "Polycentric Systems of Governance: A Theoretical Model for the Commons". In: *Policy Studies Journal* 47.4, pp. 927–952. DOI: 10.1111/psj.12212 (cit. on p. 34).
Cave, S. et al. (2019). "Motivations and Risks of Machine Ethics". In: *Proceedings of the IEEE* 107.3. Conference Name: Proceedings of the IEEE, pp. 562–574. DOI: 10.1109/JPROC.2018.2865996 (cit. on p. 19).
Cihon, Peter (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development* (cit. on pp. 19, 30).
Cihon, Peter, Matthijs M. Maas, and Luke Kemp (2020a). "Fragmentation and the Future: Investigating Architectures for International AI Governance". In: *Global Policy* 11.5, pp. 545–556. DOI: 10.1111/1758-5899.12890 (cit. on p. 21).
— (2020b). "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, pp. 228–234 (cit. on p. 21).
Council of Europe (2020). *CAHAI - Ad Hoc Committee on Artificial Intelligence (Factsheet)* (cit. on p. 25).
Dafoe, Allan (2018). *AI Governance: A Research Agenda* (cit. on p. 20).
Diab, Wael W. (2020). *Artificial Intelligence*. Geneva (cit. on p. 30).
Freedberg Jr, Sydney J. (2020). *Military AI Coalition Of 13 Countries Meets On Ethics* (cit. on p. 27).
Gasser, Urs and Virgilio A.F. Almeida (2017). "A Layered Model for AI Governance". In: *IEEE Internet Computing* 21.6, pp. 58–62. DOI: 10.1109/MIC.2017.4180835 (cit. on pp. 19, 20).
Gill, Amandeep Singh (2019). "Artificial Intelligence and International Security: The Long View". In: *Ethics & International Affairs* 33.2. Publisher: Cambridge University Press, pp. 169–179. DOI: 10.1017/S0892679419000145 (cit. on pp. 20, 25).
— (2020). *Imagining the AI future* (cit. on p. 19).
Haas, Peter M. (1992). "Introduction: epistemic communities and international policy coordination". In: *International Organization* 46.1. Publisher: Cambridge University Press, pp. 1–35. DOI: 10.1017/S0020818300001442 (cit. on p. 23).
Hao, Karen (2020). *OpenAI is giving Microsoft exclusive access to its GPT-3 language model* (cit. on p. 32).
IEEE (2021). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html, accessed 15/01/2021 (cit. on p. 30).

Imbrie, A et al. (2020). *Mainframes: a provisional analysis of rhetorical frames in AI* (cit. on p. 20).

Imbrie, Andrew et al. (2020). *Agile Alliances: How the United States and Its Allies Can Deliver a Democratic Way of AI*. Tech. rep. Center for Security and Emerging Technology, p. 84 (cit. on p. 21).

ITU (2020). *United Nations Activities on Artificial Intelligence (AI) 2020*. Tech. rep. International Telecommunication Union, p. 156 (cit. on p. 28).

Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi (2020). "Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence". In: *AI and Ethics*. DOI: 10.1007/s43681-020-00019-y (cit. on pp. 21, 24).

Jordan, Andrew et al. (2018). "Governing Climate Change Polycentrically: Setting the Scene". In: *Governing Climate Change: Polycentricity in Action?* Ed. by Andrew Jordan et al. Cambridge: Cambridge University Press, pp. 3–26. DOI: 10.1017/9781108284646.002 (cit. on p. 35).

Jordana, Jacint and Juan Carlos Triviño-Salazar (2019). "European Union Agencies: A global governance perspective". In: *Revista de Estudios Políticos* 185, pp. 169–189. DOI: 10.18042/cepc/rep.185.06 (cit. on p. 27).

Kurshan, Eren, H. Shen, and Jiahao Chen (2020). "Towards Self-Regulating AI: Challenges and Opportunities of AI Model Governance in Financial Services". In: *Proceedings of the First ACM International Conference on AI in Finance*. Association for Computing Machinery. DOI: 10.1145/3383455.3422564 (cit. on p. 19).

Larsson, Stefan (2020). "On the Governance of Artificial Intelligence through Ethics Guidelines". In: *Asian Journal of Law and Society*. Publisher: Cambridge University Press, pp. 1–15. DOI: 10.1017/als.2020.19 (cit. on p. 21).

Lepri, Bruno et al. (2018). "Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges". In: *Philosophy & Technology* 31.4, pp. 611–627. DOI: 10.1007/s13347-017-0279-x (cit. on p. 19).

MacIntyre, John, Larry Medsker, and Rachel Moriarty (2021). "Past the tipping point?" In: *AI and Ethics* 1.1, pp. 1–3. DOI: 10.1007/s43681-020-00016-1 (cit. on p. 19).

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267, pp. 1–38. DOI: 10.1016/j.artint.2018.07.007 (cit. on p. 19).

Mueller, Milton, John Mathiason, and Hans Klein (2007). "The Internet and Global Governance: Principles and Norms for a New Regime". In: *Global Governance: A Review of Multilateralism and International Organizations* 13.2, pp. 237–254. DOI: 10.1163/19426720-01302007 (cit. on p. 31).

NATO (2020). *Cooperation on Artificial Intelligence will boost security and prosperity on both sides of the Atlantic, NATO Deputy Secretary General says* (cit. on p. 27).

OBrien, Matt (2020). *US joins G7 artificial intelligence group to counter China* (cit. on p. 26).
OECD (2021). *Going Digital*. https://www.oecd.org/going-digital/ai/ accessed on 15/01/2021 (cit. on p. 29).
ÓhÉigeartaigh, Seán S. et al. (2020). "Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance". In: *Philosophy & Technology* 33.4, pp. 571–593. DOI: 10.1007/s13347-020-00402-x (cit. on pp. 19, 20).
Ostrom, Elinor (2010). "Polycentric systems for coping with collective action and global environmental change". In: *Global Environmental Change*. 20th Anniversary Special Issue 20.4, pp. 550–557. DOI: 10.1016/j.gloenvcha.2010.07.004 (cit. on p. 35).
Parson, Edward et al. (2019). "Artificial Intelligence in Strategic Context: an Introduction". In: *an Introduction*, p. 24 (cit. on pp. 20, 21).
Perspectives on Issues in AI Governance (2019). Tech. rep. Google (cit. on p. 19).
Pizzi, Mike and Mila Romanoff (2020). *Governance of AI in Global Pulse's policy work: Zooming in on Human Rights and Ethical Frameworks*. Section: News (cit. on p. 28).
Plonk, Audrey (2020). *The Global Partnership on AI takes off – at the OECD* (cit. on pp. 26, 30).
Pomares, Julia and María Belén Abdala (2020). "The G20's role and the challenge of moving beyond principles". In: *Global Solutions Journal* 5, p. 6 (cit. on pp. 21, 24).
Prakash, Abishur (2019). *The Geopolitics of Artificial Intelligence* (cit. on p. 21).
Quach, Katyanna (2019). *Nonprofit OpenAI looks at the bill to craft a Holy Grail AGI, gulps, spawns commercial arm to bag investors' mega-bucks* (cit. on p. 32).
Rasser, Martijn et al. (2019). *The American AI Century: A Blueprint for Action*. Tech. rep. (cit. on p. 21).
Reaching Critical Will (2020). *CCW Report Vol. 8, No. 2*. Tech. rep. Reaching Critical Will (cit. on p. 24).
Rotenberg, Marc (2019). *The AI Policy Sourcebook 2019*. Electronic Privacy Information Center (Epic) (cit. on p. 19).
Scherer, Matthew U. (2015). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*. SSRN Scholarly Paper ID 2609777. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.2609777 (cit. on p. 21).
Schiff, Daniel et al. (2020). "What's Next for AI Ethics, Policy, and Governance? A Global Overview". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York NY USA: ACM, pp. 153–158. DOI: 10.1145/3375627.3375804 (cit. on pp. 20, 21, 23).
Schmitt, Lewin (2020). *Global AI governance and the Brussels effect* (cit. on p. 29).
Seaman, John (2020). "China and the New Geopolitics of Technical Standardization". In: *Notes de l'Ifri*, p. 34 (cit. on p. 33).
Shi, Qian (2020). *AI Governance in 2019 - A year in review: observations from 50 global experts*. Tech. rep. Shanghai Institute for Science of Science, p. 52 (cit. on p. 20).

Stix, Charlotte and Matthijs M. Maas (2021). "Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy". In: *AI and Ethics*. DOI: 10.1007/s43681-020-00037-w (cit. on pp. 19, 21).

Thelisson, Eva, Jean-Henry Morin, and Johan Rochel (2019). "AI Governance: Digital Responsibility as a Building Block". In: p. 13 (cit. on p. 19).

Thelisson, Eva, Kirtan Padh, and L Elisa Celis (2017). "Regulatory Mechanisms and Algorithms towards Trust in AI/ML". In: p. 6 (cit. on p. 19).

Tomašev, Nenad et al. (2020). "AI for social good: unlocking the opportunity for positive impact". In: *Nature Communications* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 2468. DOI: 10.1038/s41467-020-15871-z (cit. on p. 28).

Turchin, Alexey and David Denkenberger (2020). "Classification of global catastrophic risks connected with artificial intelligence". In: *AI & SOCIETY* 35.1, pp. 147–163. DOI: 10.1007/s00146-018-0845-5 (cit. on p. 19).

United Nations (2020). *Roadmap for Digital Cooperation*. Tech. rep. United Nations Secretary-General, p. 39 (cit. on p. 27).

Veale, Michael, Max Van Kleek, and Reuben Binns (2018). "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, pp. 1–14. DOI: 10.1145/3173574.3174014 (cit. on p. 19).

Xinhua (2020). *Remarks by Chinese President Xi Jinping at 15th G20 Leaders' Summit*. http://www.xinhuanet.com/english/2020-11/21/c_139533609.htm, accessed 19/01/2021 (cit. on p. 24).

Zürn, Michael (2018). *A Theory of Global Governance*. Vol. 1. ZSCC: 0000107. Oxford University Press. DOI: 10.1093/oso/9780198819974.001.0001 (cit. on p. 27).

Chapter 2

AI AND DEMOCRACY: A BLIND SPOT IN NATIONAL AI STRATEGIES?

Abstract: Artificial intelligence (AI) technologies can have disruptive impacts on democratic societies, processes, and institutions. Whether these impacts are harmful or beneficial to democracy depends to a considerable degree on the rules and approaches with which AI is governed. As states around the world are crafting policies and regulations to shape the development and use of AI, it becomes crucial to understand how governments address its impact on democracy. This article offers two main contributions in this regard. First, based on an inductive study of national AI strategies, it proposes a novel analytical framework to systematically organise the myriad ways in which AI might impact different dimensions of democracy, for better or worse. Second, it applies this framework to a comprehensive document analysis to examine patterns in whether and how democratic governments frame these issues. To this end, the paper covers 29 national AI strategies by OECD member states from 2018 to 2023. The review shows that AI's impact on democracy is rarely covered in depth, pointing at a blind spot in democratic states' AI policies. Moreover, it unearths considerable variation in both the degree and the concrete frames with which governments address the different dimensions of democracy. The most widespread cause for concern relates to AI's impact on civil liberties, but many strategies also show awareness of the possible dangers to participation, equality, and accountability. In contrast, performance-related aspects of responsiveness and effectiveness are framed in exclusively positive terms. Based on the findings, the article offers a typology to classify government strategies based on their stances, and discusses reasons that might explain differences in their approaches.

2.1 Introduction

The rapid advances in the development and deployment of artificial intelligence (AI) technologies are expected to have a profound impact on virtually all aspects of human life. Democracy, or the health and functioning of democratic societies, is no exception (Gilardi 2022). Already in 2018, Helbing et al. (2018) sounded the alarm, asking "Will Democracy Survive Big Data and Artificial Intelligence?" Digital technologies and large online platforms have demonstrated their ability to affect political discourse and elections around the world, a trend that is likely going to be accelerated and aggravated through the proliferation of AI tools (Djeffal 2022). The generation of ever-more convincing misinformation, including so-called deepfakes, will become more accessible and harder to detect, with potentially disastrous implications for open societies (Coeckelbergh 2024). Adversaries of democratic institutions can draw on AI technology for cyberattacks on political campaigns, public organisations or critical infrastructure, further undermining trust and eroding the pillars of a healthy democratic system (A. Kaplan 2020). And at a more fundamental level, AI and related digital technologies are altering individual behaviour and social relations, e.g., by changing people's preferences and their forms of collective communication and organisation (Gallego and Kurer 2022, Risse 2023, Acemoglu and Johnson 2023).

Certainly, AI and other digital technologies also entail uncountable benefits, both for individuals and society at large. New tools may improve citizens' access to relevant information, enhance how they can express their opinions in deliberative processes, or equip policymakers with novel insights that can augment their decision-making procedures. AI may also improve the performance of public administration, rendering democratic institutions more effective and efficient.

In order to ensure that the benefits prevail over the above-mentioned risks and dangers, policymakers around the world have started to develop political and regulatory responses to the development of AI. Most of these initiatives are still at the relatively early stages of the policy cycle, namely the issue definition or agenda-setting and policy formulation stages (see Schmitt 2021). In this stage, actors are engaged in a process called policy framing, i.e., "how problems and their potential solutions are articulated and interpreted in policy debates" (Ulnicane 2022). As different actors perceive different threats or opportunities from AI, they put forward various policy frames, attempting to set the agenda and feed their respective policy preferences into the policy making process (Gilardi, Shipan, and Wüest 2021; D. Schiff 2023). Understanding these framings thus becomes a substantial research interest that has triggered a rich and growing literature.

This paper's contributions are two-fold: first, it proposes a novel framework to conceptualise the myriad ways in which the introduction of AI may impact democracy – for better or worse. Second, it applies that framework to analyse how democratic governments around the world frame issues related to the nexus of AI and democracy. Based

on a systematic and critical review of national AI strategies by 29 OECD member states from 2018 to 2023, the paper identifies several democracy-related policy frames that have emerged in the global AI policy discourse over recent years. Studying patterns and dynamics around these frames allows for answering important research questions related to how governments frame the impact of AI on democracy: Do they address it at all, and if so, which aspects do they focus on? Do they frame the impact in terms of risks or opportunities? What are the main similarities and differences across democratic governments in this regard?

Concretely, the analysis shows how the impacts of AI on democracy are barely treated in-depth in these strategies, with only a few governments dealing with the issue in-depth. Challenges to civil liberties are addressed the most, but rarely do governments go beyond rather superficial mentions of their commitment to fundamental rights. While there are notable differences between government approaches, they are, in general slightly more outspoken on the negative than on the positive consequences. However, they all acknowledge AI's potential for enhancing the performance of the public sector, which is framed exclusively in positive terms. Together with the apparent reluctance to engage with deeper political questions arising from AI, it suggests a prevalence across many democratic governments of an administrative culture that favours techno-centric approaches over deep and integrated reflections.

In the following sections, I first review relevant literature and lay out the conceptual framework. Then, I describe the systematic document review process that underpins this paper's empirical analysis, before presenting and discussing the findings and limitations.

2.2 Literature review

The impact of AI on democracy has occupied scholars and policymakers for many years (Smuha 2021; Djeffal 2022). A number of groundbreaking contributions have shed light on normative aspects (Amoore 2020), following Nemitz (2018)'s argument that "it is legitimate and necessary to ask the question how this new technology must be shaped to support the maintenance and strengthening of constitutional democracy". More recently, the field has fused the important tasks of theorising and conceptualising the ways in which AI might impact democracy with more empirically oriented discussions offering concrete evidence of these impacts (Sudmann 2019, A. Kaplan 2020, Valle-Cruz et al. 2020, Boix 2022), Christodoulou and Iordanou 2021). An especially relevant contribution in this regard comes from Smuha (2021), who puts forward a useful tripartite distinction between different types of harm that AI can cause. She highlights the importance of "societal harm" in addition to individual and collective harm, arguing that the negative societal impact is often overlooked. To illustrate this, she offers examples from the fields of equality, democracy, and the rule of law, which have also informed

this article. The specific findings and conceptual treatments of the above-mentioned contributions will be picked up again in the next section, which lays out the analytical framework.

In addition to these topical contributions, this paper – like any discussion on the role of AI in shaping political processes – is embedded in a wider academic debate on the transformations induced by digital technologies (Gilardi 2022, Acemoglu and Johnson 2023). It furthermore relates to important reflections on AI ethics (Floridi and Cowls 2019), though these are generally concerned with more universal, overarching principles that generalise beyond liberal democracies. Generally, scholarship has focused on the concern regarding potential – and realised – harms deriving from the spread of AI (Danaher 2016, Manheim and L. Kaplan 2019, Christodoulou and Iordanou 2021). However, a great number also stress the potential positive consequences that the technology may have on democracy and democratic ideals (Zarkadakēs 2020, Risse 2022).

Jungherr (2023) has put forward a useful conceptual framework for studying "the current and prospective impact of AI on democracy". For this, he distinguishes between what is today a "largely imaginary" artificial general intelligence (AGI) with superhuman capabilities, and so-called narrow AI focused on solving specific tasks. In a second step, he assesses "how AI affects different aspects of democracy, including its effects on the conditions of self-rule and people's opportunities to exercise it, equality, the institution of elections, and competition between democratic and autocratic systems of government" (ibid.). His framework is similar to the one I propose in that it identifies different levels of impact (from the individual to the societal, institutional, and systems level). However, he collapses various democratic qualities into these levels, whereas I aim to provide additional nuance by looking at different dimensions and directions in which these impacts may play out.

Such frameworks not only advance our thinking of how AI and democracy interact. They can also serve as tools to systematically analyse how politics reacts to these developments. One prominent manifestation of political response to AI, which has been the base of numerous empirical studies, are national AI strategies. While many types of actors are engaged in discussing and shaping global AI governance (see Schmitt 2021), the actions and approaches taken by national governments are undoubtedly especially relevant.

Depending on the case and the editorial and political processes that preceded the publication of national AI strategies, these documents can be seen as a reflection of wider societal inputs and deliberations, or as expressions of a government's strategic choices and preferences. In any case, they play a considerable role in shaping downstream policy outcomes, which is why it is important to understand the ways in which a given issue is framed within them (Gilardi, Shipan, and Wüest 2021). Accordingly, a number of studies has analysed the contents of various samples of AI strategies at the national (Radu 2021, Salas-Pilco 2021, Papadopoulos and Charalabidis 2020) and subnational (Liebig et al. 2022) level. Yet, these studies have placed their foci on issues other

than democracy, resulting in a persistent gap in our understanding of how governments frame this issue.

Perhaps closest to the focus of this paper, Paltieli (2022) analyses national AI strategies to contend that "these documents intervene in contemporary democratic politics by suggesting that AI can help democracies overcome some of the challenges they are facing." He identifies different kinds of imaginaries relating to democratic, sociotechnical, and data-processing aspects, which shape citizens' imaginations of what a future AI democracy could look like. While his paper contributes several important ideas, he seems to use the notions of "AI at the service of citizens" and "AI for the national interest" as equivalents to democracy, a broad-stroke definition that risks overseeing important nuances. He finds that governments, if discussing such issues at all, talk about the positive effects of improving public sector services and institutions' efficiency. Expanding his work, this article provides further concrete empirical evidence by identifying and discussing references to democratic governance found in a larger set of national AI strategies.

Political actors, such as governments, who engage in agenda-setting and problem definition may choose to focus on certain issues and highlight certain outcomes when framing AI's impact on democracy. These frames are expected to address various sets of issues and manifest competing stances towards the impact AI has on democracy (from negative threat perceptions to positive framings of benefits and opportunities). By tracing these frames through an empirical analysis of national AI strategies, this paper contributes to our understanding of the agenda-setting process of global AI policy. It identifies the ways in which policymakers around the world have defined problems related to AI and democracy, showing which of the many competing (and at times complementary) frames have resonated the most in terms of agenda-setting, and how this has affected or will affect actual policy outcomes. Moreover, the study presents theoretical contributions to the literature on AI governance. It uncovers asymmetries in the way that stakeholders frame the risks and benefits of AI, which are likely not exclusive to the discussion about democracy. Conceptualising and analysing policy discourse for other related areas might reveal similarly interesting patterns and help scholars identify distortions or blind spots. More broadly, the paper speaks to the research agenda outlined by Tallberg, Erman, et al. 2023, who argue that the "emerging AI governance should be analyzed in relation to the ideal of democracy." Subsequent research could take this further by marrying their call for normative theory on public participation and representation with the kind of empirical work conducted by D. Schiff et al. (2021), who explicitly looked for participatory processes in the design of AI ethics documents.

2.3 Analytical framework: How AI impacts democracy

This paper aims to investigate whether and how democratic governments, through their national AI strategies, address and frame the impact of AI on democracy. To put the qualitative document review on solid conceptual footing, I developed a novel analytical framework, building upon the dimensions of democratic quality put forward by Diamond and Morlino (2004). These include participation, equality, civil liberties, accountability, and responsiveness. It is worth pointing out that Diamond and Morlino (ibid.) developed their framework as a general tool for thinking about democratic quality, and did not have AI or digital technologies in mind. Still, it remains a useful blueprint that I re-fit to the AI discussion via some adjustments – updating and augmenting some of the dimensions to better reflect the observable dynamics of AI and democracy, as explained in the corresponding sections. Thus, my adapted framework serves to systematically organise the various ways – both risks and opportunities – in which AI may impact democracy's different dimensions.

As a theoretical and normative point of departure, I take Diamond and Morlino (ibid.)'s notion of a quality democracy as a system that "accords its citizens ample freedom, political equality, and control over public policies and policy makers through the legitimate and lawful functioning of stable institutions". Just like them, I distinguish between dimensions relating to the "quality of content", i.e., the fundamental principles on which democracy relies (participation, equality, civil liberties); "procedural quality", i.e. forms of citizen control over the government (vertical and diagonal accountability); and "quality of results", i.e., whether the performance and outcomes of the system satisfy citizens' expectations (responsiveness, effectiveness). The better a system performs on the various dimensions, the higher the quality of its democracy.

Diamond and Morlino (ibid.) acknowledge that the various dimensions "are so densely interactive and overlapping that it is sometimes difficult to know where one dimension ends and another begins". My changes and simplifications to the framework are thus intended to make it better fit its purpose as a heuristic tool for the subsequent document analysis. For simplicity, I dropped the "rule of law" dimension, since most of the relevant aspects are already covered by other dimensions. Then, I replaced the horizontal accountability dimension with diagonal accountability and included efficiency as an additional output-relevant component. The motivation for these choices is discussed below.

First, the framework identifies how the various aspects of democratic quality may be exposed to AI-induced influence. In doing so, the framework is reflective of various potential levers or mechanisms through which democracy may be impacted by the development and roll-out of AI. These levers are both direct (e.g., specific tools for citizens or institutions) and indirect, second-order effects that are consequences of the

spread of AI (e.g., repercussions on the quality of democratic discourse, algorithmic discrimination, legitimacy of institutions). Notably, they may affect more than one dimension simultaneously. Second, the framework discusses the potentially beneficial or harmful consequences for democracy, differentiating between risks and opportunities. These are not necessarily exclusive. Both positive and negative impacts may materialise simultaneously, albeit to different degrees, and may have mutually reinforcing dynamics. The assessment of impacts and levers includes actual as well as hypothetical – but plausible – aspects. Aiming to provide a general overview of the multiple aspects, it does not give weight to their likelihood or to the relative magnitude of impact, which would undoubtedly be a fruitful extension of this research in the future.

In the following, I discuss each of the dimensions in turn, showing how AI may or already does impact them in ways that are beneficial or detrimental to the quality of democracy, and through which levers. For ease of reference, table 2.1 provides the reader with an overview of the framework.

2.3.1 Participation

Arguably, the biggest and most imminent impact of AI relates to the participation dimension, by altering the quality of public discourse and the means for citizens' political engagement. Below, I explain what is meant by participation and why it is an essential democratic quality, before discussing the ways in which it may be strengthened or undermined by AI technologies.

Participation is an essential feature of democracy, without which the core democratic principle of popular control over government remains unattainable (see Beetham et al. 2008). To realise meaningful participation, citizens require corresponding rights, freedoms and proper means, such as access to information. As Diamond and Morlino (2004) point out, it is not enough for a regime to grant its citizens formal rights of political participation: "[A] good democracy must ensure that all citizens are in fact able to make use of these formal rights to influence the decision-making process." This requires "extensive citizen participation [...] in the life of political parties and civil society organisations, in the discussion of public policy issues, in communicating with and demanding accountability from elected representatives, in monitoring official conduct, and in direct engagement with public issues at the local level" (ibid.). Only then can participation enable a democratically legitimate authorisation of institutions through processes such as elections. [1] Here, we can already see the overlaps and interactions between dimensions: impacts on participation may also affect (vertical) accountability. Importantly, participation is also closely linked to – and contingent on – equality, as will be discussed later on.

[1] This is why I consider participation to relate predominantly to the "quality of content" of a democracy. In contrast, while Diamond and Morlino (2004) acknowledge that it speaks to both content and procedure, they place the emphasis on the latter.

	Dimension	Possible impacts and levers
Quality of content	**Participation**	Reduces/enhances quality of public discourse Distorts/facilitates citizens' opinion formation and epistemic agency
		Levers: *AI used to create disinformation (-)* *AI's potential to detect and combat disinformation (+)* *AI's role in recommender systems and filter bubbles (∼)* *AI tools for civic education and political information (+)*
	Equality	Increases/decreases exclusion and discrimination Concentrates power/democratises access to information and resources
		Levers: *Algorithmic discrimination (–)* *AI-based debiasing tools (+)* *AI-induced concentration of power (–)*
	Civil liberties	Weakens/strengthens civil liberties
		Levers: *AI used for surveillance (–)* *AI used for manipulation of individuals (–)* *AI used as a tool in political activism (+)*
Quality of procedure	**Vertical accountability**	Undermines/enhances elections Reduces/augments transparency of public sector
		Levers: *AI used for cyberattacks/-defense (∼)* *AI used for misinformation campaigns and interference (–)* *AI used to manipulate citizens (–)* *AI used to enhance public discourse (+)* *AI-based tools to control institutions and politicians (+)*
	Diagonal accountability	Weakens/strengthens free press and civil society
		Levers: *AI-induced economic pressures on media outlets (–)* *New AI-based tools to gather, analyse, and distribute information better (+)*
Quality of results	**Responsiveness**	Distorts/enhances decision-making processes Makes institutions less/more responsive to citizens
		Levers: *AI used in analytical and predictive tools (∼)* *AI-based chatbots and online services (∼)*
	Effectiveness	Reduces/enhances efficiency and effectiveness of institutions
		Levers: *AI used in analytical and predictive tools (∼)* *AI used for automatisation (∼)* *AI-based chatbots and online services (∼)*

Table 2.1: *Analytical framework of how AI may impact democracy.*

Signs in brackets indicate whether a lever's impact on that dimension is predominantly harmful (–), helpful (+) or ambivalent (∼).

In all of this, the advances of AI may affect the nature and quality of citizen participation in several ways, with potentially severe impacts on the legitimacy of democratic institutions, processes, and the wider system. This includes access to reliable information and an informed discussion of political affairs, i.e., a high-quality public sphere. By enabling and constraining the creation, dissemination, reception, and contestation of political information, AI and digital structures affect the "public arena" and shape behaviour (Jungherr 2023).

But the impacts also materialise at a more immediate, individual level. One way in which AI can bring about positive changes includes AI tools developed for purposes of civic education and to support citizens in their opinion formation process. For instance, since 2021, the technology company IBM has offered "Project Debater", described as an "AI system that can debate humans on complex topics" with the goal to "help people build persuasive arguments and make well-informed decisions" (Research 2018). As such, AI-powered tools and platforms can be used to engage citizens in democratic processes and encourage greater participation (for more on this, see Burgess 2022, Zarkadakēs 2020). Moreover, AI-driven chatbots may answer questions about the democratic process and offer information about voting, political parties' positions, and civic engagement (Schneier, Farrell, and Sanders 2023). Advanced election recommendation systems may use AI to build "digital twins" of citizens and identify their optimal vote choices that best align with their preferences (Heesen et al. 2021). Naturally, these tools also open the doors for misuse and manipulation. Proper regulatory safeguards as well as broad civic education will be indispensable to ensure that citizens can adequately use them to enhance their participation in the political process.

AI-driven recommender systems have also been the focus of concern for another level at which AI may impact participation. By controlling what users get to see online, AI systems are essential to the quality of public discourse, potentially undermining citizens' ability to make informed decisions and participate effectively in democratic processes (A. Kaplan 2020). One such phenomenon concerns epistemic bubbles (often referred to as filter bubbles or echo chambers), which expose users predominantly to information that reinforces their existing political views. While the exact effects are contested, the overarching concern is that it causes citizens to increasingly gravitate towards ideological extremes, decreasing possibilities for compromise due to mistrust of institutions, the media, and citizens with different political views (Barberá 2020). Moreover, social media algorithms that prioritise user engagement may be more likely to promote fabricated stories than the truth (Giansiracusa 2021). So as AI accelerates the spread of disinformation in various ways, it contributes to political polarization and the erosion of societal consensus on key political facts (see Acemoğlu 2021; Acemoğlu 2023).

And beyond distribution, AI is also increasingly used in the creation of misinformation. The latest generation of generative AI floods the internet with photo-realistic images, plausible-sounding text bites, and increasingly even synthetic audio and video content that non-experts cannot easily identify as inauthentic (so-called "deepfakes").

While their precise political and psychological effects are still being investigated (see Walker, D. S. Schiff, and K. J. Schiff 2024), early studies produce mixed evidence, with Dobber et al. (2021) finding some negative effects on voters' perceptions of politicians, especially in connection with microtargeting practices, whereas the results of an experiment by Hameleers, Meer, and Dobber (2022) suggest that "the strong societal concerns on deepfakes' destabilizing impact on democracy are not completely justified." However, the technology is improving quickly, and in a more recent experiment, Hameleers, Meer, and Dobber (2024) already find that "participants have a hard time distinguishing a deepfake from a related authentic video". The authors further speak of a "climate of factual relativism arguably contributes to overall distrust and doubt among news users, who are not always able or willing to accurately separate facts from fiction." The prevailing concern is that deepfakes and other AI-powered technologies open the gates for coordinated disinformation campaigns at ever larger scales and with ever more convincing means. It also lowers the technical barriers for malign individuals or groups, which gain unprecedented means for cheaply and quickly creating compelling, but inauthentic, content. This risks polluting the information environment, as social media platforms and journalistic outlets, let alone ordinary citizens, struggle to verify the authenticity and accuracy of images, videos, audio, or any other reported intelligence. All of this pushes societies towards a dangerous path of post-factualism (Ferretti 2021). Without reliable access to accurate and truthful information, the ideal of an enlightened democratic discourse gets ever harder to attain.

However, AI may also have positive effects, facilitating high-quality online deliberation and public participation in decision-making processes. For instance, AI tools can be used to detect and suppress misinformation or other harmful content (Giansiracusa 2021). And recommender algorithms may be tooled to expose citizens to conflicting viewpoints and explain their underlying justifications, thus increasing mutual understanding (Barberá 2020). Or by filtering out or downplaying potentially conflicting or otherwise harmful issues, they may decrease their salience.[2]

In sum, while there are clear risks of AI to "diminish the epistemic agency of citizens and thereby undermine the relevant kind of political agency in democracy" (Coeckelbergh 2022), it may also enhance participation. Beyond the levers considered so far, there are concomitant issues that affect participation in more indirect ways. The two most important of these – the potentially exclusive aspects of AI and its challenges to fundamental rights and freedoms – are discussed thoroughly in the following sections.

[2] Whether such practices would be ethical and democratic and who would control the underlying decisions requires an entirely different, highly fascinating debate. However, that would go beyond the scope of the article, for which it suffices to say that it is imaginable that AI would be used for such purposes.

2.3.2 Equality

Indeed, issues around equality of citizens are closely related to participation, which requires formal and actual equality in order to be meaningful. On the individual level, equality in this sense means that each citizen has the same weight in the political process. This formal precondition is in practice largely distorted by unequal distributions of wealth, status, political access, and education, as well as other exclusionary factors. As discussed below, AI may have additional impacts in this regard. On the collective level, democratic equality means that the political system is not dominated by certain powerful interest groups. This stands at odds with the practical reality of lobbying and other ways through which large corporations and resourceful groups can exercise "more power to shape public debate and preferences and to determine the choice of leaders and policies" (Diamond and Morlino 2004). And again, developments in the AI realm will likely play a role in shaping these aspects of equality further.

There are multiple ways in which AI may plausibly affect power balances and how inclusive or exclusive the democratic system becomes. On the individual level, AI can exacerbate existing inequalities and biases. Numerous AI algorithms have been shown to perpetuate historical discrimination against certain groups or individuals, particularly those from marginalised communities (Mitchell et al. 2021; Ferrer et al. 2021). The exclusionary consequences of such discrimination affect marginalised individuals and groups in multiple harmful ways, of which the implications on democracy are only one of many. A major concern in this regard is that meaningful participation is hampered because citizens suffering from discrimination may lack the necessary resources or avenues to engage in politics. This, in turn, will also have implications for how representative democratic institutions are.

From an economic perspective, Acemoğlu (2021) warns that "automation shifts the balance of power away from labour towards capital, and this can have far-ranging implications on the functioning of democratic institutions." Concretely, AI's impact on labour markets may further undermine employees' bargaining power, which weakens their means of representation and influence (Gallego and Kurer 2022; Jungherr 2023). It also poses threats to welfare systems which are generally built on workers' contributions (Greve 2019) and highly exposed to structural employment shifts induced by technological change such as AI.

Moreover, we have witnessed over recent years an enormous agglomeration of power in the hands of a few large technology companies and individuals (Verdegem 2022). This concentration is highly problematic for democracy, as critical decisions affecting all of humankind are vested in the hands of a few powerful corporations lacking any democratic legitimacy or effective control (Zuboff 2019; Acemoglu and Johnson 2023). Moreover, it gives these individual actors an outsized influence over policymaking, whether through traditional forms of lobbying or by using their control over technologies and digital ecosystems for political means (Farrow 2023). It is still unclear whether the lat-

est developments in AI technology mark a shift towards a more open and democratic ecosystem or the opposite (Widder, West, and Whittaker 2023), but it is something that governments and societies around the world should be closely watching.

But AI's trajectory does not necessarily have to result in more inequality and harms to democracy – it can also be a force for democratisation (Acemoglu and Johnson 2023). Indeed, AI may also serve to decrease discrimination and enhance equality within a society (Y.-T. Lin, Hung, and Huang 2021). For instance, for AI used in hiring, Chamorro-Premuzic (2019) proposes to train it "to ignore people's gender and focus only on the relevant signals of talent or potential." Such positive scenarios are much less talked about, with most observers focusing on the ways in which AI may increase discrimination. The ample evidence of algorithmic discrimination and biased AI seems to justify this one-sided focus. However, for a constructive, solution-oriented discussion, one should also consider the ways in which AI may serve the opposite effect. Indeed, it could even be seen as a moral obligation to shape the development of technology in such a way that it increases, rather than decreases, inclusiveness and equality.

Moreover, AI-powered tools and technologies can help to make the democratic process more accessible to individuals with disabilities or other barriers to equal participation. For example, AI-powered voice recognition and natural language processing technologies can be used to make online content available to individuals with visual or hearing impairments.

2.3.3 Civil liberties

Another key dimension of democracy relates to freedoms guaranteed through certain individual rights. In this regard, Diamond and Morlino (2004) distinguish between political, civil, and social (or socioeconomic) rights. While all of them are undoubtedly important, my framework focuses on civil rights only, due to conceptual and empirical considerations. Political rights, including the right to vote, to stand for office, to campaign, and to organise political parties, are rarely directly affected by AI. Moreover, conceptually, I consider these rights to be subsumed in the dimensions of participation and vertical accountability (elections). Social or socioeconomic rights, such as the right to private property or to collective bargaining, may more plausibly be affected by AI. However, I again choose not to list them separately, as I conceive them to be conceptually embedded in the dimension of equality. This leaves me with (fundamental) civil rights and liberties, such as freedom of thought, expression, and information; or the freedom of assembly, association, and organisation (ibid.). Conceptually, these freedoms are both a necessity for and a consequence of a healthy democracy, and can be heavily affected by AI developments, as argued below.

When it comes to freedom of thought, expression, and information, the possible impact of AI technologies is closely related to the quality of the democratic discourse mentioned before. These freedoms can be curtailed by AI systems that contribute to the

pollution of the information space, that suppress or augment certain voices, or that even manipulate citizens subconsciously. Moreover, AI technologies are extremely prone to invading privacy and can be used to infer citizens' preferences (Kosinski 2021) or predict their political behaviour (Argyle et al. 2023). Such knowledge might allow political opponents, oppressive state apparatuses, or other malicious actors to take preemptive measures, thus directly or subversively undermining the right of assembly, association, and organisation. In other words, AI-powered surveillance technologies give governments and political campaigns unprecedented abilities to monitor and manipulate the electorate.

Contrary to such a bleak outlook, AI-enhanced means of collective organisation could help politically active citizens to utilise their freedoms more readily. The broad and cheap availability of AI tools may democratise access to information and resources that are relevant for meaningfully expressing one's voice. This would allow previously marginalised individuals and communities to communicate their concerns and preferences more effectively. For instance, translation tools can help non-native speakers comprehend and express themselves more easily. AI tools for content creation may substitute part of the work of communication agencies, which can be prohibitively expensive and thus discourage people from pursuing political goals.

AI may also serve to protect citizens by limiting the state's violations of civil liberties. For instance, it has been argued that the use of AI in law enforcement may – if introduced properly – actually help to "deracialise policing", rather than augment existing discrimination (Capers 2016). Importantly, both positive and negative dynamics may be occurring simultaneously. However, the positive aspects in this regard are less evident and may weigh much less than the above-mentioned harms.

Evidently, this dimension is closely related to equality. This applies all the more to other fundamental rights and freedoms, such as those related to non-discrimination and equal opportunities. While AI is unlikely to result in any immediate change in the formal rights prescribed to citizens in democracies, the introduction of opaque and complex algorithmic systems may make it much harder for citizens and public authorities to properly enforce and protect these rights (Busuioc 2021; Izdebski 2023). For instance, the use of AI in law enforcement and the judiciary has repeatedly been found to be biased against certain groups, thereby undermining their rights to non-discrimination and equal treatment. Sadly, though, such bias has preceded AI systems, which are reflective of them. However, such structural bias can be increasingly hard – if not impossible – to detect for an individual when it is caused by opaque AI systems. This is especially true if the affected individual lacks the knowledge and resources necessary to understand the functioning of such AI systems and subsequently challenge the outcomes, be they court decisions or police actions.

These arguments tie in with some of the levers concerning accountability, discussed in the next section.

2.3.4 Vertical and diagonal accountability

AI may pose a threat to democracy by potentially weakening the ability of the political system to provide checks and balances on executive power. Accountability – in a democratic context – is "the obligation of elected political leaders to answer for their political decisions when asked by voters or constitutional bodies" (Diamond and Morlino 2004). It thus comprises at least two directions: vertical accountability, which describes the link between citizens and leaders, or principals and agents; and horizontal accountability, which relates to the division of powers across government pillars, and the fact that democratic officeholders and institutions must answer also to other institutions (O'Donnell 1998). Recent scholarship has proposed the inclusion of another link, called "diagonal accountability" (Lührmann, Marquardt, and Mechkova 2020), which refers to the role of non-state actors in ensuring accountability. This includes actors outside of formal political institutions, such as civil society organisations, a free press, and politically active citizens, who can provide and amplify information about the government. Importantly, these diagonal accountability mechanisms depend heavily on the other two lines of accountability to function effectively, and are themselves conditional upon other constitutive elements of democracy, such as freedom of expression.

For my framework, I choose to exclude horizontal accountability, as there are no obvious potential or observable direct impacts of AI on the division of power across institutions. The document review process (see next section) also did not reveal any instances of governments discussing AI's impact on horizontal accountability. However, it is worth noting that AI may become a factor in horizontal accountability under conditions of extreme political polarization. The rise of "constitutional hardball" strategies (i.e., when political actors depart from any forbearance in the pursuit of power, even if it means pushing constitutional limits and violating democratic norms, see Levitsky and Ziblatt 2018) renders plausible that different government branches might use AI to attack and discredit each other. For instance, in the US, former President Donald Trump's repeated attacks on the legal system risk diminishing the courts' ability to hold executive government to account (Nelson and Gibson 2019.[3] Future research endeavours should pay close attention to such dynamics, as this paper only focuses on vertical and diagonal accountability.

The most prominent feature of vertical accountability in democracies is the holding of regular, free, and fair elections. In states that rely on digital forms of voting, AI may pose its most immediate threat to the integrity of elections by providing attackers with sophisticated new tools to disrupt or manipulate the voting process (Manheim and L. Kaplan 2019). However, it may just as well be used to safeguard election systems, enhancing their cybersecurity and resilience against such threats. The second, more complex and concerning mechanism in which AI may impact elections relates to the

[3] I am grateful to my reviewers for pointing this out.

quality of the public discourse, already discussed previously.

For both vertical and diagonal accountability to work properly, the quality of public deliberation is crucial. Extending the previous line of thought about the potential deterioration or improvements to the overall quality of the democratic discourse, accountability may be helped or hindered. Fuelled by AI, fake news and polarization may upend traditional mechanisms for accountability such as established political norms that render certain behaviour as (un-)acceptable. Politicians who were repeatedly and severely lying generally used to face electoral or legal consequences. In a setting where the information environment is polluted by fake news, citizens increasingly struggle to distinguish lies from truth, and accountability mechanisms fail.

Jungherr (2023) puts forward another mechanism through which AI may reduce the legitimacy of elections. It builds on Przeworski (1991)'s concept of "institutionalised uncertainty" – the fact that elections allow for the possibility of unforeseen political outcomes and eventual change in government – as an essential condition of elections. As Jungherr (2023) explains, the predictive power of AI might "threaten to offset this perceived uncertainty of who will lose and who will win elections."

Besides, the introduction of AI into public institutions may be detrimental for accountability in other ways. Algorithm-based decision-making introduces intransparency and thereby undermines the legitimacy of key political processes (Danaher 2016). For example, if AI is used to make decisions in areas such as law enforcement or social welfare, it may be difficult for citizens to understand how those decisions are made and to hold public officials accountable for their actions (Izdebski 2023; Busuioc 2021). The problem starts with the general opaqueness of many AI systems due to their often complex black-box architecture and often nontransparent deployment. Moreover, "AI algorithms have also been found to get caught in negative feedback loops that are difficult to spot, break out of, and/ or self-correct", potentially leading to "self-fulfilling prophecies" (Busuioc 2021). For example, if an AI system designed to detect social security fraud is trained on historically biased data, it will disproportionally target discriminated groups, further amplifying that same discrimination while hiding it behind the seemingly objective functioning of an algorithm. For the affected groups, it will be very hard to challenge the model assumptions and underlying training data, or the decisions, without proper safeguards or supervision.

These are some of the levers in which AI may harm accountability. However, if AI is harvested in a way that enhances the information landscape and improves the overall quality of democratic discourse, accountability mechanisms may even be strengthened. Indeed, AI-powered tools may plausibly offer new ways of generating transparency. Today's complex democratic systems produce a vast amount of information. Simultaneously, numerous decisions are delegated to closed-door bodies and specialised agencies, often without undergoing much public scrutiny (Mounk 2018). AI tools that allow the automated tracking of legislative and other regulatory procedures, that summarise legal developments and highlight the most important aspects to the user, may all play a role

in improving the transparency of democratic processes. Also, AI-powered tools can be used to monitor government spending and detect instances of corruption, which can help hold public officials and institutions accountable. New tools may allow citizens to better track the performance and behaviour of their elected politicians or democratic institutions at large, thus enhancing their position when it comes to the exercise of accountability. A playful illustration in this regard was a project by Belgian artist Dries Depoorter, who developed an AI tool to detect parliamentarians who used their phones during plenary sessions. Screenshots of these politicians were then automatically uploaded to social media (Depoorter 2023).

The availability of such new tools enhancing the information landscape connects directly to diagonal accountability, or the ability of the press and civil society to exercise control public officials and institutions. Jungherr (2023) identifies several ways in which AI negatively affects the economic conditions of news production, putting additional pressure on journalists and media outlets. This ranges from the move towards automated content production built on generative AI to how AI-driven search and recommender systems channel audiences to their news sources, rendering a "decline in monetization opportunities of news" (ibid.) with an especially strong impact on smaller, local news outlets.

In contrast, B. Lin and Lewis (2022) discuss how AI in journalism may be beneficial for democracy, focusing on its role in information gathering, selection and production, and distribution and consumption. They explain how and why "journalistic AI" could "work in the service of accuracy, accessibility, diversity, relevance, and timeliness." In other words, AI may be used to strengthen independent investigations and critical reporting of government actions by the press, thereby enhancing diagonal accountability mechanisms. Similarly, it may provide civil society actors and activists with new means to gather, analyse, and distribute evidence of political malpractice, thus strengthening their role as watchdogs and societal corrective players.

2.3.5 Responsiveness

Meaningful participation, equality, and civil liberties can be seen as necessary preconditions or inputs for a well-functioning democracy. In turn, forms of accountability refer to procedural qualities. Finally, one ought to consider the system's output performance or "quality of results", for it "influences the degree to which citizens will be satisfied with the performance of democracy and view it as legitimate" (Diamond and Morlino 2004). A major part of this relates to the capacity (and willingness) of officials and institutions to properly aggregate citizens' preferences and translate them into policies. Thus, performance is largely conditioned by the responsiveness and quality of decision-making. I add to this another output-relevant component, which is acknowledged but not explicitly addressed by Diamond and Morlino (ibid.): effectiveness and efficiency. This will be discussed in the following section.

The first component, responsiveness of institutions and quality of decision-making, relates to democratic institutions at the highest level (e.g., the degree to which citizens can make their voices heard in the decision-making of governments and whether the ultimate decisions resemble good choices) as well as to lower levels of public administration (e.g., whether local government is responsive to the needs and demands of ordinary citizens). The promises of AI in this regard are clear: It may provide institutions, elected officials, and policymakers with novel means to take into account citizens' concerns and preferences during decision-making processes. Moreover, AI may provide additional information and serve as an enhancer of the quality of decisions, if implemented well (see, for instance, Valle-Cruz et al. (2020)). Singapore is widely considered a trailblazer in this regard, featuring inter alia an AI-powered chatbot that allows residents to report municipal issues directly via WhatsApp and Telegram. On the flip side, though, a hasty and careless introduction of AI into democratic institutions and public administrations may actually hurt their standing, if citizens feel that decisions are arbitrary or reflecting bad judgement.

2.3.6 Effectiveness and efficiency

The second component of performance, effectiveness and efficiency, is closely related but focuses less on the alignment between citizens' preferences and outputs, and more on the operational aspects of how outputs are generated. The way in which AI can affect processes and procedures of governments and public administration is certainly not unique to democracies. Still, I deem it relevant to incorporate these aspects into the analysis because of the significant indirect impact of effectiveness and efficiency on democratic support. While empirical evidence suggests only weak links between government effectiveness and abstract support for democracy (Claassen and Magalhães 2022), ineffective governance does reduce government approval (Romero, Magaloni, and Díaz-Cayeros 2016; Eichengreen, Saka, and Aksoy 2022). Accordingly, politicians and citizens worry about trade-offs between efficiency and democracy (Wilson 1901; Gaenslen 1980). Democratic governments may be more efficient than other systems due to better accountability mechanisms (Adam, Delis, and Kammas 2011), but collective decision-making processes are never perfect. If outcomes are excessively misaligned with the expectations of citizens, calls for alternatives to democracy grow louder. This may incentivise policymakers to consider how they could use AI to improve the efficiency of public administrations and state institutions. Insofar as it is implemented carefully and properly, this may be a good thing for democracy, as increased performance enhances output legitimacy and support for the system. However, it also entails a range of risks, including to democratic qualities such as the above-mentioned fundamental rights.

There are many ways in which AI can be used to streamline public administration and improve the efficiency and effectiveness of government services. AI-powered chatbots can be used to provide information and support to citizens faster and cheaper. For

instance, since 2016, the Australian Taxation Office (ATO) offers a chatbot for public inquiries, saving clients and staff time and money (Henman 2020). Thus, satisfaction by citizens and staff is coupled with economic advantages. AI can help to reduce the costs of public administration by accelerating procedures, automating or augmenting repetitive tasks, and improving the accuracy of decision-making processes. This could lower the burden on tax-paying citizens and free up public resources to be used elsewhere in the provision of public goods and services. Overall, these developments should serve to increase citizens' satisfaction with public administration and democratic institutions and, therefore, their legitimacy.

However, the introduction of AI into public administration and democratic processes and institutions may also have detrimental impacts on their performance. As this would decrease citizens' satisfaction, it would by extension also pose a serious risk to the legitimacy of democracy at large. Conceivable ways in which this could happen include, but are not limited to: faulty output of algorithmic decision-making tools, bad political choices informed by inaccurate or poorly designed AI tools, misalignment between AI suggestions and citizens' actual preferences, technical hurdles in the implementation of AI systems (e.g., cybersecurity or privacy issues or the creation of unmanageable complexity).

2.3.7 Interactions between dimensions

As the discussion has shown, several levers affect multiple dimensions at once, and at times it is challenging to pinpoint which dimension is the most adequate to capture a given phenomenon. For instance, AI-powered misinformation clearly diminishes citizens' means of participating in an informed democratic debate. But in doing so, it also severely undermines channels of accountability. Or consider AI tools in public administration that follow open data principles. Ideally, this should not only augment the performance of the authorities, but also serve to increase transparency. Additionally, if this data is sourced by investigative journalists or data-driven democracy AI tools, it may also enhance citizens' resources to get informed about crucial democracy-related issues. These examples illustrate the often cross-cutting effects that AI may have across the various facets of the framework. They further highlight the blurry nature of the labels, reminding us that the framework is primarily a heuristic tool to serve analytical purposes, not a fully accurate and precise representation of reality.

These dimensions and the outlined ways in which they are or might be impacted by AI are by no means exhaustive. There are several other direct and indirect ways in which AI can impact democracy. And there might be other dimensions or aspects related to democracy in which the advantageous and disadvantageous effects of AI may materialise. For instance, economic development, cultural norms, or international relations may shape and be shaped by AI, with important implications for the quality of democracy. AI-triggered economic transformations (such as growing wage inequalities, job

replacement, etc.) could plausibly fuel rising social tensions and democratic instability (Acemoğlu 2023).

Moreover, I purposefully restricted the discussion to scenarios that are realistically achievable with today's technologies or plausibly within reach over the short- to medium-term. While it is important to also consider the long-term impacts of AI, these are by definition more speculative and therefore less productive for an analytical framework designed to investigate existing national AI strategies, whose definitions of AI are overwhelmingly anchored in the short- and medium-term.

2.4 Systematic review of AI policy documents: How governments discuss AI and democracy

The empirical analysis of this paper is based on a systematic review of selected national AI strategies or comparable documents. Restricting the data set to texts released by national governments of member states of the OECD was motivated by two considerations. First of all, since the analysis is focused on AI's impact on democracy, it seems especially relevant to study what democratic governments have to say about the issue. Membership of the OECD implies that countries are generally seen as democratic. Moreover, the sample consists of relatively comparable countries in terms of economic development and governance structures. Potential shortcomings of the sample restriction are discussed in the section on limitations (3.7.6).

2.4.1 Methodology

I employ a (light) quantitative content analysis based on (thorough) qualitative reading of the documents as it allows for a systematic examination and interpretation of textual data. Qualitative document analysis, which enables researchers to uncover patterns, themes, and meanings within a corpus of documents (Elo and Kyngäs 2008) has several advantages. For instance, data is stable and – unlike interviews or field observations – it cannot be influenced by the researcher in the collection process (Morgan 2022). However, the method also bears several limitations, such as limited access to documents or an inherent subjectivity in the reading and interpretation of texts. In general, this approach is particularly suited for exploring complex phenomena, such as the impact of AI on democracy, because of its flexibility and ability to capture the nuances and context-dependent nature of political text (Graber and Smith 2005). Furthermore, qualitative content analysis allows for iterative and reflexive engagement with the data, aligning with the iterative nature of qualitative inquiry (Schreier 2012). The process of coding, categorization, and interpretation allows to systematically analyse the content of AI documents while remaining open to emergent themes and unexpected insights (Hsieh and Shannon 2005).

The document collection was done in August 2023 via the OECD.ai Policy Observatory, a repository of official government policies and strategies. To ensure that the sample was complete and accurate, I carried out additional desk research. The full number of OECD member states is 38, but not all member states have a publicly available strategy, if any at all. This brings the sample of strategies to be considered for this paper down to 32 strategies, of which 24 are available in English. I could analyse a further five in their original languages (i.e., in French, German, and Spanish). The remaining three (Iceland, Latvia, and Slovenia) had to be discarded because no authoritative translation was available. Annex 2.8 provides a full list of all 29 national AI strategies that were included in the analysis, as well as some additional considerations regarding the selection procedure.

Each selected strategy has been read carefully in its entirety, enabling the identification and extraction of passages that resonate with the key elements of the analytical framework outlined in section 2.3 or any other, more general, democracy-relevant points. In this, I followed an inductive trajectory, designed to ensure the completeness of the framework by facilitating the discovery of new insights and perspectives that might enrich our understanding of the complex interplay between AI strategies and democratic considerations. When the framework had been revised based on new frames unearthed in the process, documents were analysed again to account for the changes. While each strategy has been read at least once, the iterative nature of the process meant that some strategies or parts thereof had been read multiple times, to account for the evolving framework.

For a passage to be considered relevant for the final analysis, it had to meet at least one of the following two criteria: 1) it had to explicitly speak to one or more of the dimensions of democracy contained in the analytical framework; 2) it had to explicitly address one of the levers in which AI may affect one or more of the dimensions, and situate this in a democracy-relevant content. Thus, a passage speaking about the integrity of elections in the age of AI would immediately classify as relevant because it meets the first condition. A passage speaking about algorithmic discrimination would only classify as meeting the second condition if it relates this to, for instance, political representation, participation, and equality. However, if it relates this to, for instance, business or consumer aspects, it would not classify because it does not fully meet the second condition. Similarly, a passage speaking about human autonomy or potential manipulation only classifies if it explicitly or implicitly relates this to the political system.

In the end, I identified 365 relevant passages across all 29 documents. The count of passages per document has a median of 11 and a mean of 12.6. The minimum number of passages in a document is 2 and the maximum is 38, demonstrating the broad variation in attention paid to the impact of AI and democracy, as unpacked further below. I carefully inspected and coded each of the 365 relevant passages, based on the framework's elements and framings they speak to. Notably, a passage could be assigned to

more than one dimension. Moreover, to fully capture the framing, I also coded the stance associated with a passage. If a passage indicated concern, mentioned risks, or reflected upon negative developments that are or might be occurring due to AI, it has been labelled "negative". If it expressed optimism or hope, mentioned advantages, or reflected upon positive developments that are or might be occurring due to AI, it has been labelled "positive". If a passage hinted at both risks and opportunities, it has been marked as "ambivalent", but was coded for both the positive and negative frame. This only affected a handful of passages and helped simplify the subsequent presentation and interpretation of the findings. Passages that did not indicate a positive or negative framing were coded as "neutral". Taken together, each combination of dimension and stance is a possible frame. With a total of seven dimensions and three different stances (positive, neutral, negative), this results in 21 possible frames that may be presented by democratic governments in their national AI strategies.

In the following, I first present extracts from the review, which illustrate the ways in which democratic governments have addressed the different dimensions and put forward different frames. I then proceed by discussing several high-level findings that emerge from the analysis.

2.4.2 Variation across dimensions

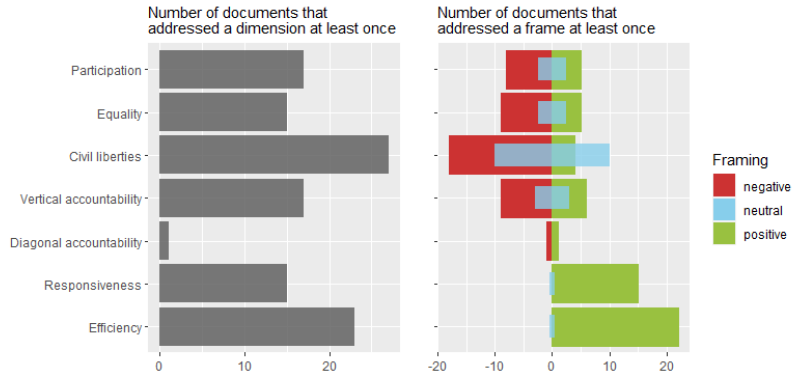
Reporting the results of the qualitative analysis of national AI strategies, figure 2.1 displays information on two important counts. The left-hand plot shows how many documents mention a given dimension at least once, whereas the right-hand plot shows the breakdown by frames (i.e., dimensions combined with stance). Below, I discuss each dimension in turn and cite examples of the various frames detected in government strategies.

Participation

More than half of the documents (17/29) refer to questions relating to AI affecting citizens' capacities to participate in democratic deliberations and processes. Among them, only eight express concern about the potential harms, mostly referring to its potential for polluting or distorting democratic discourse. For instance, the German strategy mentions AI's potential for "changing the social interaction patterns and debates that are required to ensure our democratic order" (p. 42). While this is a rather abstract definition, others worry more concretely about influencing campaigns and manipulation, e.g., through deepfakes (Austria, Australia, United Kingdom).

Five strategies mention the potential of AI for improving participation. For instance, Finland points out AI's potential for "increasing citizens' participation in decision-making and democratic processes" (p. 38), while Belgium highlights the opportunities of using AI to combat disinformation. A strategy which explicitly mentions opportu-

Figure 2.1: *Count of documents addressing a given dimension/frame at least once*



Note: The sum of counts on the right-hand panel may exceed the left-hand panel, which collapses all frames of a dimension into a single count, and can therefore not exceed 29 (the total number of strategies).

nities for collective participation is the Spanish strategy, highlighting the possibility of "creating bottom-up democratic, participatory processes with the use of new technologies" (p. 70).

Equality

Equality is addressed in 15 out of 29 documents, slightly less than participation but with a similar distribution of the positive (5) and negative (9) frames. A number of states are concerned about discrimination and equal treatment, though only a few make the link to democracy and political processes. Some stand out for explicitly mentioning the risk that AI may concentrate power in the hands of a few big technology firms, with Germany warning that this "must not lead to scientists and civil society becoming ever more dependent on obtaining financing from these companies" (p. 43). And Japan cautions that "[t]he use of AI should not generate a situation where wealth and social influence are unfairly biased towards certain stakeholders" (p. 10). Curiously, those very same governments also often see AI as a potential tool in fighting for more equality. The German strategy calls for "innovative applications that support self-determination, social inclusion, cultural participation" (p. 39). Others highlight the opportunities of AI technology for combatting discriminatory treatment of citizens by public officers (Poland, p. 26) and making public services more accessible (Finland, p. 123).

As an interesting side note, very few strategies consider another dynamic linked to equality, but not directly to democracy (and therefore not a focus of this paper): the global inequality between rich, developed nations and what is often called the "Global South" when it comes to developing and deploying AI, and shaping global AI policy. Switzerland is a remarkable exception in this regard, as it calls for global inclusion into

AI policy debates, explicitly advocating for developing countries (p. 5). The only other country to acknowledge this problem is Chile, which is geographically placed in the Global South, but economically and politically part of the club of developed nations.

Civil liberties

Civil liberties are a broad, fundamental dimension and it is therefore hardly surprising that more governments make mention of it (27/29). This is coupled with a very large number of neutral references (found in 20 of the strategies). Notably, the number of documents mentioning opportunities (4) is almost the same as for the equality dimension, whereas there is a much higher number of documents speaking about the risks (18). To be fair, many of these negative frames are rather inconcrete statements, mainly reaffirming the government's commitment to fundamental rights which may be threatened by AI (e.g., Ireland, p. 15 and p. 40). Amongst specific civil liberties and rights, privacy is the one most frequently mentioned in regards to AI. This makes sense, given the data-hungry logic of AI solutions, and their demonstrated risk in exposing or revealing critical personal information. Some strategies explicitly mention the importance of upholding the right to free thought and speech in the age of AI. For instance, Estonia reaffirms the "protection from direct and indirect interference, deception, and manipulation" (p. 42). Chile, the Netherlands, and the United States all express concern about AI's potentially chilling impact on freedom of speech.

Meanwhile, there are a few who also see AI's potential for protecting or enhancing civil liberties. For instance, Germany wants to fund "innovative applications that support self-determination, social inclusion, cultural participation and the protection of privacy" (p. 39). Ireland makes reference to the potential of AI to improve fairness in the judicial system – though it also warns about the associated dangers.

Vertical accountability

The vertical accountability dimension is addressed in 17 out of 29 strategies. Six states seem to appreciate the potential of AI for enhancing accountability in democratic affairs. These include links to transparency in democratic processes and public administration. For instance, the French strategy proposes to use AI "to help individuals better understand administrative rules and how they apply to their personal situations." The Spanish strategy also addresses these frames, calling for a register of automated systems within government administrations to document them (p. 69). The Spanish strategy doubles down on these aspects, highlighting how AI can help "improve the transparency and disclosure of public activity" (p. 57) and add transparency to decisions (p. 58). Others mention the potential of using AI to combat disinformation, which of course also plays into the participation dimension. But when linked to elections, it directly affects the vertical accountability mechanisms of a system.

This is also the area where most concerns are expressed within this dimension. Nine states highlight the potential of AI to be misused for manipulation and misinformation with the potential of affecting elections or the democratic process in general. Only a few, however, consider how the introduction of AI into decision-making procedures may harm transparency and, by extension, accountability. The Italian strategy stands out in this regard, with very clear language on the importance of accountability and the "legal liability upstream of certain decisions or results" that were based on AI algorithms (p. 51). Austria is another case that focuses on transparency and traceability of AI-based decisions in public administration, suggesting the introduction of control mechanisms (p. 57). And the Finnish strategy writes that the "increasing use of AI and automated decision-making in public services raises also major ethical issues and challenges relating to, for example, transparency and supervision" (p. 108).

Diagonal accountability

Diagonal accountability, i.e., the possibilities of non-state actors such as the press and civil society to report on the state and pressure for change in cases of misbehaviour, is addressed directly only in the German strategy. When discussing the danger of AI concentrating ever more power in the hands of a few big tech companies, the government commits to "[enabling] scientists and civil society to provide independent and skills-based contributions to this important public debate" (p. 43). It proceeds by discussing the possibilities and limitations of AI concerning freedom of opinion, information, the media, and art – highlighting the importance of the press and culture for "free, individual and public opinion-formation" (p. 44).

It is remarkable that no other strategy made reference to diagonal accountability. Encouragingly, many strategies engaged with civil society in the process of formulating their AI policies. Still, they should not overlook the potential risks that AI-driven developments may pose to the press (see Jungherr (2023)'s discussion on the "Economics of News" for more). A more in-depth discussion of these dynamics would ideally also unearth other, positive ways in which AI may actually strengthen the press and civil society, thus contributing to diagonal accountability and overall democratic quality.

Responsiveness

Of all the framework's dimensions of AI's impact on democracy, frames relating to performance were by far the most common and the most positive. In fact, not one of the 15 documents speaking about responsiveness or of the 23 documents speaking about efficiency ever related to this in a negative way.

Instead, states regularly use their AI strategies to frame the technology as a means to enhance the responsiveness of public administration and quality of decision-making. Mostly, this relates to the way in which citizens can communicate and interact with the

state, whether public services can be better targeted to citizens' needs, and whether AI can aid or augment decision-making at various stages of government and the public sector. While some speak in rather abstract terms of "new channels for citizen attention" (Chile, p. 10, own translation) or AI's potential to enhance "citizens' interactions with public authorities by providing improved, tailored public services" (Luxembourg, p. 16), others are more specific. For instance, the Estonian strategy envisions AI to "help to make better political decisions based on data, e.g., by using machine learning and methods of artificial intelligence for detecting patterns in information" (p. 18). A number of documents appreciate the use of AI for evaluative, predictive, and diagnostic purposes, or even for the "[d]evelopment of complex modelling systems to simulate decision-making situations" (Hungary, p. 38).

Efficiency

Many strategies mention also the potential of AI to lead to more efficient and effective institutions. The Spanish strategy is particularly explicit in this regard, stating that "AI enables public agencies to become more effective and efficient, and to enhance their relations with society" (p. 58). Similarly, the Danish strategy emphasises the potential "to support faster and more efficient case processing" (p. 10). The Danish government explicitly commits to using AI to "improve the quality of citizen service centres, make problem-solving more efficient, and increase confidence in the public administration" (p. 56).

It is worth noting that most of the instances in which governments highlight the opportunities of AI for improving institutional efficiency are not directly linked to democracy as such. Instead, they refer to public administrations and political institutions in a general manner, and imply that their effectiveness and proper functioning is a goal in itself. A few remarkable exceptions exist in this regard. For instance, the Italian strategy spells out the link between institutional performance and democracy in very clear terms, pointing out that as "services become more efficient, relations with citizens are improved and the level of trust in institutions is increased" (p. 54).

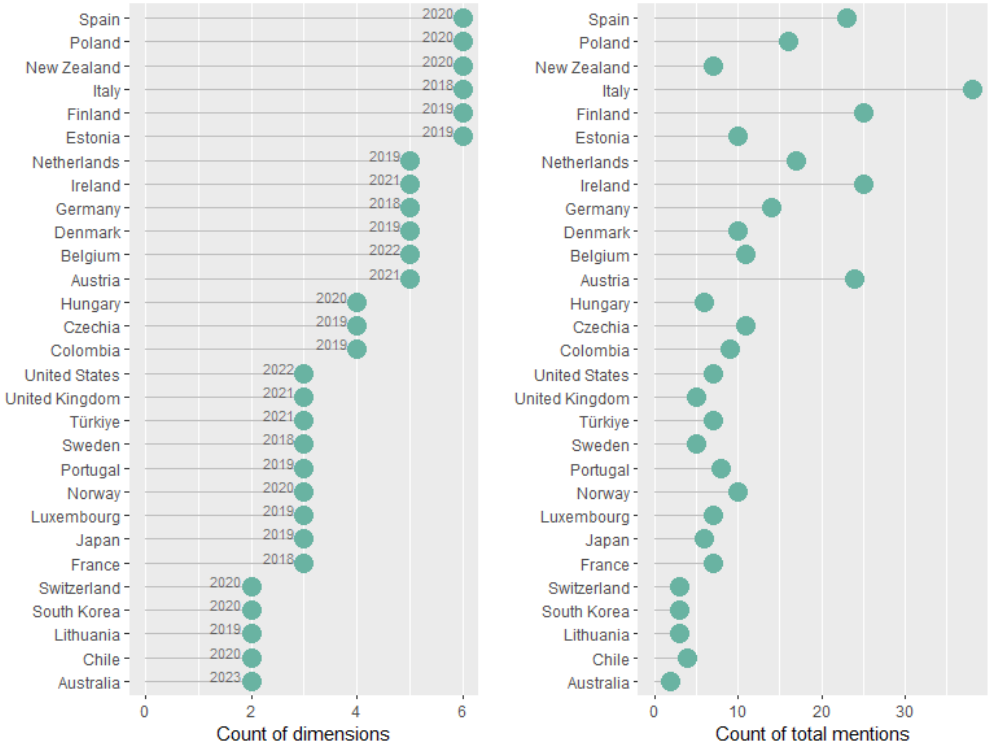
2.4.3 Variation across countries

In addition to high variation across different dimensions and associated framings, there is also high variation in how extensively government strategies deal with AI's impact on democracy. Figure 2.2 illustrates this, ranking countries in order of the number of dimensions they covered within their strategies (left-hand plot) and contrasting this with the total count of democracy-related references detected per document (right-hand plot).

On the one end, countries such as Spain, Poland, New Zealand, Italy, and Finland cover a very broad range of issues, checking five out of the framework's seven dimen-

sions. On the other end, there are countries with very low engagement, such as Australia, Chile, or Lithuania. Notably, all strategies address at least two of the dimensions. The depth of engagement, illustrated by the total count of references, is not perfectly correlated with the breadth, i.e., the number of different dimensions addressed.[4] This is mostly explained by the varying document length. For instance, New Zealand's AI Principles document covers several themes in just four pages, so naturally each theme is not addressed more than once. The Italian strategy, on the other hand, spans 40 pages, so its authors can refer to similar themes repeatedly throughout the text. By and large, however, breadth and depth largely map onto each other, showing a high correlation of 0.74.

Figure 2.2: *Count of dimensions addressed at least once within a strategy (left-hand plot) and total mentions of democracy-related themes per document (right-hand plot).*



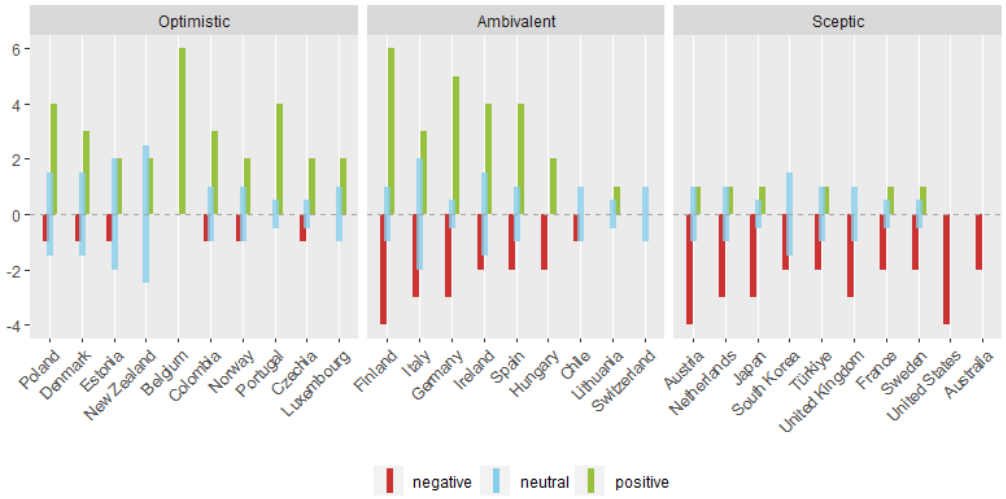
Note: The maximum count of dimensions that could be observed is seven.

Countries not only differ in the scope of their engagement with AI's impact on

[4] Analysing AI documents based on their breadth and depth is inspired by D. Schiff et al. 2021, who did a similar exercise to illustrate how different sectors engage with AI ethics.

democracy. They also frame differently the various positive and negative consequences, as illustrated by figure 2.3. It plots the number of different frames per country, grouping them into one of three larger camps: the "optimistic" countries are those where governments frame the nexus of AI and democracy predominantly positively; the opposite cases are classified as "sceptic" countries. The remaining countries, in which neither framing dominates, are labelled as "ambivalent".

Figure 2.3: *Count of frames addressed at least once within a strategy, by groups.*



Each camp of countries is of roughly similar size. The ambivalent camp includes numerous cases where strategies engaged with the nexus of AI and democracy either very little – and hence it did not make sense to assign them to either end of the scale – or a lot, but in a rather balanced fashion. It stands out that the optimistic camp is comprised of comparably smaller countries, whereas the democratic AI powerhouses UK and US can both be found in the sceptic camp.

Lastly, it is important to reiterate that these labels refer specifically to how countries frame AI's impact on democracy – not how they assess its overall potential to do harm or good. Some strategies are positive throughout, whereas others are generally optimistic, but make several negative statements when it comes to democracy. The UK is a good example in this regard, as it clearly chooses to embrace AI as a transformative technology with many opportunities, while also being among the most vocal nations when it comes to the harms.[5] The degree of threat perception is probably also a reflec-

[5]The discussion of harms by the UK is not limited to democracy-related aspects. It is also the only country which explicitly introduces longtermist discussions around so-called existential or catastrophic risk, artificial general intelligence (AGI), and alignment (p. 9 and 17). This illustrates the influence of the controversial Effective Altruism movement on UK government policy (Clarke 2023).

tion of how convinced policymakers are of the transformative (or incremental) changes that AI will usher in. Someone who views it as a powerful, singular force for profound changes will appreciate the many opportunities and the necessity of harnessing these. Thus, the produced strategy may be very positive of AI overall. However, when it comes to specific topics such as democracy, the very same beliefs may inspire caution or outright alarm. And on the other end of the spectrum, policymakers who do not buy into the promises associated with AI will likely see less reason to ponder about its potential harms to democracy.

2.5 Discussion: a blind spot in national AI strategies

The systematic review revealed several noteworthy observations regarding the way that OECD member states framed the possible impacts of AI on democracy. One is the overall low level of attention to the issue. The other is the high focus on performance or output-related aspects, and their exclusively positive framing. In the following, I discuss these overarching patterns and suggest some possible explanations. Lastly, I acknowledge some limitations of this study and caveats for interpreting the results.

Sociopolitical issues are largely neglected in many of the strategies, which instead tend to focus much more on industry- and business-related aspects such as investment, technical aspects such as data infrastructures, and accompanying measures such as research and education. When societal issues are introduced, they are often limited to nods to the global AI ethics discussions, sustainability and climate change, or the labour market. Rarely do they address AI's possible impacts on democracy.

Alarmingly, a majority of strategies do not explicitly address many of AI's possible impacts on democracy at all, hinting at a huge blind spot by democratic governments around the world. If anything, these documents refer vaguely to the public interest or threats to fundamental values. And even in those strategies where some mentions can be found, these do not receive a lot of attention compared to many other societal issues.

Moreover, even those strategies that address a broader range of democracy-related aspects rarely go into depth. Instead, the mentions are often generic and superficial, giving the sensation that this was more of a checkbox exercise rather than a profound engagement with the issues.

There are, however, noteworthy exceptions of countries that demonstrate a higher willingness to engage with these discussions and where policymakers seemingly have a stronger notion of its importance. Examples of strategies that go deeper are the Austrian and Finnish strategies, which not only cover a broad range of dimensions, but also do so extensively. Moreover, the qualitative reading of their strategies revealed that they discuss AI's impact on democracy in much more specific terms than many of the other strategies do.

Related to the tendency of strategies to focus on technical aspects seems to be the

strong attention to AI's potential in enhancing outputs, which often reflects a rather technocentric approach. Across all strategies, the performance-related dimensions were covered the most, and exclusively in positive terms. Framing the introduction of better technology as an enormous opportunity makes sense to generate societal buy-in, signal the need for reform to public servants, and perhaps even spur investments by private actors into relevant gov-tech solutions. However, it seems risky to completely ignore the potential ways in which the introduction of AI technology may also have negative implications. An informed and balanced debate should at least acknowledge the risks, even if the bottom line emphasises the opportunities more strongly.

More gravely, though, it seems that the focus on performance distracts policymakers from the other ways in which AI might impact democracy. The analysis found more frames related to efficiency and responsiveness than to all the other dimensions combined. This pattern is remarkably stable across countries. Only in three cases is efficiency not the most discussed dimension (Germany, Finland, and the Netherlands).

It would be helpful to understand what factors are driving the degree to which governments are discussing AI's impact on democracy, and in which frames. To this end, a number of possible explanations are explored below, not to offer conclusive evidence but to stimulate interpretation of the observed patterns.

A key factor to determine the outlook of national AI strategies could be the government's ideology. Studying the AI Act negotiations in the European Parliament, Chiappetta (2023) finds that "each of the main political parties delineate distinct party-driven perspectives". This begs the question whether such unique party positions can indeed be identified and generalised to explain policy choices by governments based on their political affiliation. Left-leaning governments might be more prone to speak about civil liberties and equality, whereas right-leaning governments may focus on performance. Populists may opt to ignore the civil liberties and diagonal accountability dimensions, and therefore display a more narrow breadth of dimensions addressed in their strategies. However, when controlling for such simplified political orientations after assigning government into corresponding camps (left- vs right-leaning, populist vs mainstream), no systematic relationship between government ideology and outcomes could be observed. This is probably due to the novelty of AI in political debates and the lack of clear positions that can be distinguished across party lines or ideological camps (Ulnicane and Erkkilä 2023). The development of more concrete AI policies, in combination with the rising salience of AI, may likely change this and force political actors to take a stance. Moreover, once AI's impacts will be felt at scale, i.e., on the labour market, its politicisation will certainly follow suit.

Other explanations regarding the detected patterns relate to possible temporal trends. For instance, one may assume that governments have initially overlooked the potential of AI in impacting democracy, but over time learned from experts, experience, and each other. Curiously, though, temporal trends point to the opposite direction. Early strategies from 2018 and 2019 – which make up 14 of the 29 strategies – are somewhat more

likely to cover a wide range of dimensions and frames than the 15 strategies published in 2020 and thereafter. Several explanations for this are plausible, which are by no means mutually exclusive. Perhaps policymakers were initially eager to present a comprehensive, fully-fledged account of AI and society, thus looking into as many areas as possible. For later publications, policymakers had time to learn from their predecessors and the ongoing political debate on AI, resulting in more focused and narrower treatments. Or, after the initial hype and almost hysteria following the Brexit referendum and Donald Trump's election as US president in 2016, policymakers were initially more attuned to the risks. This perceived threat immediacy was less present to policymakers in later years, which might have interpreted the absence of high-level events as evidence that these concerns are inflated or misplaced and do not need to be engaged with by government strategies. Another, more cynical, argument could pinpoint this decreasing engagement with democracy to the rising influence of technology corporations. As the market for AI exploded in recent years, so did their lobbying efforts to steer AI policy into more business-friendly waters. Dwelling on AI's impact on democracy might be seen as counterproductive to their interests, thus pushing governments to adapt a more narrow, investment- and innovation-minded perspective in their AI strategies.

These explanations stand at odds, however, with a second temporal trend that can be observed from the analysis. Notably, earlier strategies were more likely to be positive in their coverage of democracy-relevant issues, whereas most of the more recent strategies fall into the ambivalent or sceptic camps. This suggests a completely different learning pattern in how policymakers perceive and frame AI's impact on democracy. Over time, they become more focused and more critical, demonstrating a higher willingness to engage with the harms associated with the technology.

This would offer some relief to those who are concerned that democratic governments are not sufficiently attuned to the dangers that AI might pose to the political system. It also fits with more recent public statements by leading politicians, who have repeatedly voiced their concern in this regard. Future research could investigate these statements and other sources, thus offering additional evidence to advance this discussion. It would also be valuable to expand the analysis to other democratic states that are not part of the OECD, as well as to contrast these findings with the strategies and approaches of non-democratic states. And lastly, as AI politics is evolving, it will also become increasingly relevant to study the actual policy outcomes of all these debates.

In sum, there is ample room for future research that disentangles and explains the variation across government approaches to AI's impact on democracy. Beyond pointing at possible explanations, this discussion has highlighted that my present study is better suited to identify aggregate trends rather than variation within the sample. This is as much due to the small sample size as to the many unobserved factors affecting the data generation process, i.e., how documents are authored within government systems.

2.6 Limitations and caveats

One of the primary limitations of this study is the inherent subjectivity involved in human coding of documents. This process introduced subjectivity and potential bias, as many decisions required difficult judgement calls. Different coders may interpret and categorise information differently, which could impact the consistency and reliability of the results.

The analytical framework that I proposed and on which the document review is built may be incomplete. To account for this, I designed it during an iterative process in which I repeatedly read the strategies and reworked the framework's categories until it satisfyingly covered all relevant aspects brought up in the strategies. Dimensions from the original list proposed by Diamond and Morlino (2004), which I dropped, also did not emerge during the review process, though there is reason to believe that horizontal accountability may become increasingly important in this regard, as mentioned in the introduction. Moreover, one may argue that a discussion on what member states are potentially missing in their strategies should not be limited to aspects they (or at least some of them) actually talk about. However, grounding the framework in actually mentioned frames also binds it to reality and plausible, realistic concerns.

Another limitation relates to the case selection. My study focused on analysing AI strategies from OECD member states. This case selection was deliberate and driven by the need for comparability among relatively similar democratic governments with established AI policies. However, this choice has the regrettable limitation of excluding voices and perspectives from the Global South. These regions have unique challenges and approaches to AI policy, and their marginal role in the formulation of global AI policy has also been problematised in this very paper. Still, factors such as language barriers and varying degrees of AI policy development caused me to opt against widening my study at this point.

Moreover, my study focused on national AI strategies as the primary data source. While these documents provide valuable insights into governments' formal positions and strategies regarding AI and democracy, my approach naturally has limitations. By concentrating only on official policy documents, I left aside other relevant data sources, such as speeches, op-eds, interviews with policymakers, and public statements. These more immediate expressions could provide valuable additional insights into policymakers' concerns and evolving perspectives on AI's impact on democracy. However, opening the analysis to such data sources would have made it highly difficult to restrict its scope to a manageable and comparable level.

When it comes to interpreting the findings, and especially the absence of certain frames, another important caveat applies: the absence of evidence for policymakers considering a given frame may not necessarily imply evidence of them being unaware of unconcerned. There may be other reasons for which a certain frame is not included in the final document. These reasons may be due to editorial and political choices, if the au-

thors of a strategy deem an issue to be out of the intended scope of the strategy. Or they may be strategic, if mentioning an issue would go against the underlying interests of the government. For instance, technocratic preferences for a more automated, efficient public sector may discourage policymakers from discussing the potential downsides of introducing AI in administrations. This analysis can of course only consider what has been written in these strategies, whereas it remains in the realm of speculation – and ideally future research – to explore what has not been written and why. Few strategies are as transparent about policymakers' considerations and weightings of options as the Italian strategy, which ponders: "Can the public decision-maker transfer his political responsibility to an AI system that does not respond to a clear principle of representation?" (p. 36). In sum, the present analysis is probably better suited to show and discuss aggregate trends rather than variation within the sample, due to its small size and the many unobserved factors affecting the data generation.

Furthermore, the analysis is agnostic to the fact that different elements of the framework may be more or less likely to materialise, and that their associated impacts may differ in magnitude. Giving weights to different aspects may be a challenging enterprise, but would undoubtedly be a valuable extension of this research in the future.

More generally speaking, restricting the analysis to democracy-related matters excluded other effects of AI, positive or harmful, which a holistic, comprehensive assessment might also wish to include. Ultimately, though, there is a trade-off to be made between scope and accuracy. Furthermore, extending the assessment to a broader range of issues would immediately raise additional normative challenges of how impacts in different areas are weighted against each other or whether some fundamentals such as democracy ought indeed to be non-negotiable.

By design, this study is focused on framing and policy formulation rather than implementation. While national AI strategies offer a glimpse into governments' intended approaches, they do not necessarily reflect the actual impact on democracy. Policymaking is a complex process, and policies may not always translate into concrete actions or outcomes as intended. Future research should explore the implementation and effectiveness of AI policies in safeguarding democratic principles. Moreover, we urgently need more research, societal debate, and political solutions around the potential harms that AI may pose to the various dimensions of democracy as outlined in this paper.

2.7 Conclusion

This article has developed a novel analytical framework to disentangle the various ways in which AI impacts democracy. It thus provides a systematic approach to organising and categorising the potential impacts, helping to facilitate nuanced analyses of AI's effects on democratic processes, institutions, and actors. Applying the framework to a qualitative document analysis of 29 national AI strategies by OECD member states

from 2018 to 2023, the article has detected which dimensions and frames are addressed by governments. This generates valuable insights into their level of concern and respective focus areas. Taken together, the proposed analytical framework and the results of the document analysis can serve as a useful tool for scholars and policymakers to understand the diverse ways in which AI can influence democratic societies. By examining the competing issue-definitions and frames, the article's empirical analysis sheds light on states' preferences and approaches to regulating AI in the context of democracy. This understanding of governmental perspectives can provide valuable context for scholars and policymakers seeking to engage in discussions and debates surrounding AI governance.

Concretely, the findings suggest that AI's impact on democracy is largely overlooked in these strategies, with only a few governments dealing with the issue in-depth. Challenges to civil liberties are most frequently addressed, but even then governments tend to stick to rather superficial mentions of their commitment to fundamental rights. They also rarely consider the interaction of different dimensions, or how other sociopolitical impacts (e.g., on the labour market) may affect democracy at large. In general, governments are slightly more outspoken on the negative than on the positive consequences. Nevertheless, all of them acknowledge AI's potential for enhancing the public sector's performance, an aspect which is framed exclusively in positive terms. Together with the apparent reluctance to engage with deeper political questions arising from AI, it suggests a prevalence across many democratic governments of an administrative culture that favours techno-centric approaches over deep and integrated reflections. Still, there are notable differences between government approaches. The presence of different frames allows to classify strategies into optimistic, ambivalent, and sceptic groups of roughly equal size, with the AI powerhouses UK and US falling into the latter category. While government ideology does not seem to be systematically related to these different approaches, the analysis hints at a slight temporal trend. Strategies published more recently are less comprehensive and more negative when it comes to AI's impact on democracy.

Taken together, the findings suggest that democratic governments around the world need to better anticipate AI's impact on democracy. They should develop and pursue policies that harness opportunities while safeguarding against the risks. Many of the issues identified in the framework and by various strategies require dedicated efforts to better educate citizens, both on the specific consequences of AI and digital technologies as well as on democracy more broadly. Governments should continue to fund research and civil society initiatives in these regards. Other issues such as algorithmic discrimination, AI safety, or privacy-related matters can best be addressed through targeted regulation on AI technology, its developers and users. In this regard, existing efforts such as the EU's AI Act will be influential. As policymakers around the world negotiate the rules for AI, they should make sure that they are conducive for and protective of democratic principles and practices. Ultimately, the review of national strategies highlighted

an overarching willingness to deploy AI in the public sector in order to boost the performance of institutions. In doing so, the public sector has a chance and responsibility to establish best practices and set high-quality standards which can spill over into the private sector. Governments should equip their institutions with sufficient resources and a strong legal basis to roll out AI in a responsible manner.

The analysis exposed an asymmetry in the way that stakeholders think about the harms and benefits of AI. This asymmetry is by no means exclusive to AI's impact on democracy – ambivalent consequences can be expected in many other aspects, but are not always adequately addressed in policy discussions. For some aspects, risks are downplayed while opportunities get highlighted, as was found to be the case for efficiency in administrations (see also Toll et al. 2020 for how the Swedish public sector's policy discourse on AI "may be overly optimistic"). For other aspects, there may be undue emphasis on the harms and little regard to AI's potential. A point in case would be the discourse on privacy, which has traditionally focused heavily on the risks deriving from AI models' churning through personal data. Much less space has been given to the potentially privacy-enhancing features of AI technology (see Liu et al. 2018; Ustundag Soykan et al. 2022). Another point in case is the debate on the political impact of AI-powered disinformation through deepfakes. While these have typically been framed as threats to democracy (Vaccari and Chadwick 2020), more recent research suggests that "evidence on a strong persuasive advantage of deepfakes compared to other forms of disinformation and authentic content in the political realm is lacking" (Hameleers, Meer, and Dobber 2024). However, the assessment of whether such distortions or asymmetries exist largely relies on subjective judgement by observers and is rarely quantified. This article presents a remedy by putting forward a simple and straightforward research design based on a framework coupled with qualitative document analysis that generates evidence to substantiate claims of asymmetric policy debate.

Overall, the article's contributions enrich the academic discourse on the relationship between AI and democracy, providing valuable insights for scholars, policymakers, and practitioners engaged in AI governance and the preservation of democratic principles in the digital age. It lays the groundwork for further research on the complex interplay between AI technologies and democratic societies and hopefully stimulates discussions on the crucial role of AI in shaping the future of democracy. The article also shows the divergence in how seemingly like-minded governments approach important questions of AI governance. It thus confirms previous studies that identified similar differences, e.g., across EU member states (Djeffal, Siewert, and Wurster 2022) or across sectors (Tallberg, Lundgren, and Geith 2023). While fragmentation of approaches around the world may be an obstacle to successful international cooperation, such a competition of ideas and policies may also serve as a useful and efficient way to identify the most robust and reliable responses to the risks and challenges posed by AI, regarding democracy and beyond.

Bibliography

Acemoglu, Daron and Simon Johnson (2023). *Power and progress: our thousand-year struggle over technology and prosperity*. London: Basic Books (cit. on pp. 42, 44, 51, 52).

Acemoğlu, Daron (2021). *Dangers of unregulated artificial intelligence* (cit. on pp. 49, 51).

— (2023). "Harms of AI". In: *The Oxford Handbook of AI Governance*. Ed. by Justin B. Bullock et al. 1st ed. Oxford University Press, C65P1–C65N5 (cit. on pp. 49, 59).

Adam, Antonis, Manthos D. Delis, and Pantelis Kammas (2011). "Are democratic governments more efficient?" In: *European Journal of Political Economy* 27.1, pp. 75–86. DOI: 10.1016/j.ejpoleco.2010.04.004 (cit. on p. 57).

Amoore, Louise (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press (cit. on p. 43).

Argyle, Lisa P. et al. (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3. Publisher: Cambridge University Press, pp. 337–351. DOI: 10.1017/pan.2023.2 (cit. on p. 53).

Barberá, Pablo (2020). "Social Media, Echo Chambers, and Political Polarization". In: *Social Media and Democracy: The State of the Field, Prospects for Reform*. Ed. by Joshua A. Tucker and Nathaniel Persily. SSRC Anxieties of Democracy. Cambridge: Cambridge University Press, pp. 34–55 (cit. on pp. 49, 50).

Beetham, David et al. (2008). *Assessing the quality of democracy: a practical guide*. Stockholm: Internat. IDEA (cit. on p. 47).

Boix, Carles (2022). "AI and the Economic and Informational Foundations of Democracy". In: *The Oxford Handbook of AI Governance*. Ed. by Justin B. Bullock et al. 1st ed. Oxford University Press. DOI: 10.1093/oxfordhb/9780197579329.013.64 (cit. on p. 43).

Burgess, Paul (2022). "Algorithmic augmentation of democracy: considering whether technology can enhance the concepts of democracy and the rule of law through four hypotheticals". In: *AI & SOCIETY* 37.1, pp. 97–112. DOI: 10.1007/s00146-021-01170-8 (cit. on p. 49).

Busuioc, Madalina (2021). "Accountable Artificial Intelligence: Holding Algorithms to Account". In: *Public Administration Review* 81.5, pp. 825–836. DOI: 10.1111/puar.13293 (cit. on pp. 53, 55).

Capers, I. Bennett (2016). "Race, Policing, and Technology". In: *North Carolina Law Review* 95, p. 1241 (cit. on p. 53).

Claassen, Christopher and Pedro C. Magalhães (2022). "Effective Government and Evaluations of Democracy". In: *Comparative Political Studies* 55.5. Publisher: SAGE Publications Inc, pp. 869–894. DOI: 10.1177/00104140211036042 (cit. on p. 57).

Clarke, Laurie (2023). "How Silicon Valley doomers are shaping Rishi Sunak's AI plans". In: *POLITICO* (cit. on p. 67).

Coeckelbergh, Mark (2022). "Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence". In: *AI and Ethics*. DOI: 10.1007/s43681-022-00239-4 (cit. on p. 50).

— (2024). *Why AI undermines democracy and what to do about it*. Cambridge Hoboken, NJ: Polity (cit. on p. 42).

Chamorro-Premuzic, Tomas (2019). "Will AI Reduce Gender Bias in Hiring?" In: *Harvard Business Review*. Section: Hiring and recruitment (cit. on p. 52).

Chiappetta, Allessia (2023). "Navigating the AI frontier: European parliamentary insights on bias and regulation, preceding the AI Act". In: *Internet Policy Review* 12.4 (cit. on p. 69).

Christodoulou, Eleni and Kalypso Iordanou (2021). "Democracy Under Attack: Challenges of Addressing Ethical Issues of AI and Big Data for More Democratic Digital Media and Societies". In: *Frontiers in Political Science* 3 (cit. on pp. 43, 44).

Danaher, John (2016). "The Threat of Algocracy: Reality, Resistance and Accommodation". In: *Philosophy and Technology* 29.3. Publisher: Springer Verlag, pp. 245–268. DOI: 10.1007/s13347-015-0211-1 (cit. on pp. 44, 55).

Depoorter, Dries (2023). *The Flemish Scrollers, 2021-2023* (cit. on p. 56).

Diamond, Larry Jay and Leonardo Morlino (2004). "The Quality of Democracy: An Overview". In: *Journal of Democracy* 15.4, pp. 20–31. DOI: 10.1353/jod.2004.0060 (cit. on pp. 46, 47, 51, 52, 54, 56, 71).

Djeffal, Christian (2022). "Democracy, AI Regulation and the Draft EU AI Act". In: *Turkish Policy Quarterly* (cit. on pp. 42, 43).

Djeffal, Christian, Markus B. Siewert, and Stefan Wurster (2022). "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies". In: *Journal of European Public Policy* 29.11, pp. 1799–1821. DOI: 10.1080/13501763.2022.2094987 (cit. on p. 74).

Dobber, Tom et al. (2021). "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" In: *The International Journal of Press/Politics* 26.1. Publisher: SAGE Publications Inc, pp. 69–91. DOI: 10.1177/1940161220944364 (cit. on p. 50).

Eichengreen, Barry, Orkun Saka, and Cevat Aksoy (2022). "The political scar of epidemics". In: *The Economic Journal* (cit. on p. 57).
Elo, Satu and Helvi Kyngäs (2008). "The qualitative content analysis process". In: *Journal of Advanced Nursing* 62.1, pp. 107–115. DOI: 10.1111/j.1365-2648.2007.04569.x (cit. on p. 59).
Farrow, Ronan (2023). "Elon Musk's Shadow Rule". In: *The New Yorker* (cit. on p. 51).
Ferrer, Xavier et al. (2021). "Bias and Discrimination in AI: A Cross-Disciplinary Perspective". In: *IEEE Technology and Society Magazine* 40.2, pp. 72–80. DOI: 10.1109/MTS.2021.3056293 (cit. on p. 51).
Ferretti, Maria Paola (2021). "Post-Factualism, Political Communication and the Role of Citizens". In: *Virtues, Democracy, and Online Media*. Num Pages: 17. Routledge (cit. on p. 50).
Floridi, Luciano and Josh Cowls (2019). "A Unified Framework of Five Principles for AI in Society". In: *Harvard Data Science Review*. DOI: 10.1162/99608f92.8cd550d1 (cit. on p. 44).
Gaenslen, Fritz (1980). "Democracy vs. Efficiency: Some Arguments from the Small Group". In: *Political Psychology* 2.1. Publisher: [International Society of Political Psychology, Wiley], pp. 15–29. DOI: 10.2307/3790968 (cit. on p. 57).
Gallego, Aina and Thomas Kurer (2022). "Automation, Digitalization, and Artificial Intelligence in the Workplace: Implications for Political Behavior". In: *Annual Review of Political Science* 25.1, pp. 463–484. DOI: 10.1146/annurev-polisci-051120-104535 (cit. on pp. 42, 51).
Giansiracusa, Noah (2021). *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*. Berkeley, CA: Apress (cit. on pp. 49, 50).
Gilardi, Fabrizio (2022). *Digital Technology, Politics, and Policy-Making*. 1st ed. Cambridge University Press. DOI: 10.1017/9781108887304 (cit. on pp. 42, 44).
Gilardi, Fabrizio, Charles R. Shipan, and Bruno Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". In: *American Journal of Political Science* 65.1, pp. 21–35. DOI: 10.1111/ajps.12521 (cit. on pp. 42, 44).
Graber, Doris A. and James M. Smith (2005). "Political Communication Faces the 21st Century". In: *Journal of Communication* 55.3, pp. 479–507. DOI: 10.1111/j.1460-2466.2005.tb02682.x (cit. on p. 59).
Greve, Bent (2019). "The digital economy and the future of European welfare states". In: *International Social Security Review* 72.3, pp. 79–94. DOI: 10.1111/issr.12214 (cit. on p. 51).
Hameleers, Michael, Toni G. L. A. van der Meer, and Tom Dobber (2022). "You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media". In: *Social Media + Society* 8.3. Publisher: SAGE Publications Ltd, p. 20563051221116346. DOI: 10.1177/20563051221116346 (cit. on p. 50).

Hameleers, Michael, Toni G. L. A. van der Meer, and Tom Dobber (2024). "They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes". In: *European Journal of Communication* 39.1. Publisher: SAGE Publications Ltd, pp. 56–70. DOI: 10.1177/02673231231184703 (cit. on pp. 50, 74).

Heesen, Jessica et al. (2021). *KI-Systeme und die individuelle Wahlentscheidung Chancen und Herausforderungen für die Demokratie*. Tech. rep. München: Lernende Systeme (cit. on p. 49).

Helbing, Dirk et al. (2018). "Will Democracy Survive Big Data and Artificial Intelligence?" In: *Towards Digital Enlightenment*. Ed. by Dirk Helbing. Cham: Springer International Publishing, pp. 73–98. DOI: 10.1007/978-3-319-90869-4_7 (cit. on p. 42).

Henman, Paul (2020). "Improving public services using artificial intelligence: possibilities, pitfalls, governance". In: *Asia Pacific Journal of Public Administration* 42.4, pp. 209–221. DOI: 10.1080/23276665.2020.1816188 (cit. on p. 58).

Hsieh, Hsiu-Fang and Sarah E. Shannon (2005). "Three Approaches to Qualitative Content Analysis". In: *Qualitative Health Research* 15.9, pp. 1277–1288. DOI: 10.1177/1049732305276687 (cit. on p. 59).

Izdebski, Krzysztof (2023). *The Digital Battlefield for Democratic Principles*. Tech. rep. National Endowment for Democracy (cit. on pp. 53, 55).

Jungherr, Andreas (2023). "Artificial Intelligence and Democracy: A Conceptual Framework". In: *Social Media + Society* 9.3. Publisher: SAGE Publications Ltd. DOI: 10.1177/20563051231186353 (cit. on pp. 44, 49, 51, 55, 56, 64).

Kaplan, Andreas (2020). "Artificial Intelligence, Social Media, and Fake News: Is this the End of Democracy?" In: *Digital Transformation in Media & Society*. Ed. by Ayşen Akkor Gül, Yıldız Ertürk, and Paul Elmer. Istanbul University Press. DOI: 10.26650/B/SS07.2020.013 (cit. on pp. 42, 43, 49).

Kosinski, Michal (2021). "Facial recognition technology can expose political orientation from naturalistic facial images". In: *Scientific Reports* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 100. DOI: 10.1038/s41598-020-79310-1 (cit. on p. 53).

Levitsky, Steven and Daniel Ziblatt (2018). *How democracies die*. First edition. New York: Crown (cit. on p. 54).

Liebig, Laura et al. (2022). "Subnational AI policy: shaping AI in a multi-level governance system". In: *AI & SOCIETY*. DOI: 10.1007/s00146-022-01561-5 (cit. on p. 44).

Lin, Bibo and Seth C. Lewis (2022). "The One Thing Journalistic AI Just Might Do for Democracy". In: *Digital Journalism* 10.10, pp. 1627–1649. DOI: 10.1080/21670811.2022.2084131 (cit. on p. 56).

Lin, Ying-Tung, Tzu-Wei Hung, and Linus Ta-Lun Huang (2021). "Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias". In: *Philosophy & Technology* 34.1, pp. 65–90. DOI: 10.1007/s13347-020-00406-7 (cit. on p. 52).

Liu, Bo et al. (2018). "Using Adversarial Noises to Protect Privacy in Deep Learning Era". In: *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6. DOI: 10.1109/GLOCOM.2018.8647189 (cit. on p. 74).

Lührmann, Anna, Kyle L. Marquardt, and Valeriya Mechkova (2020). "Constraining Governments: New Indices of Vertical, Horizontal, and Diagonal Accountability". In: *American Political Science Review* 114.3. Publisher: Cambridge University Press, pp. 811–820. DOI: 10.1017/S0003055420000222 (cit. on p. 54).

Manheim, Karl M. and Lyric Kaplan (2019). "Artificial Intelligence: Risks to Privacy and Democracy". In: *Yale Journal of Law & Technology* 21, pp. 106–188 (cit. on pp. 44, 54).

Mitchell, Shira et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163. DOI: 10.1146/annurev-statistics-042720-125902 (cit. on p. 51).

Morgan, Hani (2022). "Conducting a Qualitative Document Analysis". In: *The Qualitative Report* 27.1, pp. 64–77. DOI: 10.46743/2160-3715/2022.5044 (cit. on p. 59).

Mounk, Yascha (2018). "The Undemocratic Dilemma". In: *Journal of Democracy* 29.2, pp. 98–112. DOI: 10.1353/jod.2018.0030 (cit. on p. 55).

Nelson, Michael J. and James L. Gibson (2019). "How Does Hyperpoliticized Rhetoric Affect the US Supreme Court's Legitimacy?" In: *The Journal of Politics* 81.4, pp. 1512–1516. DOI: 10.1086/704701 (cit. on p. 54).

Nemitz, Paul (2018). "Constitutional democracy and technology in the age of artificial intelligence". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133. Publisher: Royal Society, p. 20180089. DOI: 10.1098/rsta.2018.0089 (cit. on p. 43).

O'Donnell, Guillermo A (1998). "Horizontal Accountability in New Democracies". In: *Journal of Democracy* 9.3. Publisher: Johns Hopkins University Press, pp. 112–126. DOI: 10.1353/jod.1998.0051 (cit. on p. 54).

Paltieli, Guy (2022). "The political imaginary of National AI Strategies". In: *AI & SOCIETY* 37.4, pp. 1613–1624. DOI: 10.1007/s00146-021-01258-1 (cit. on p. 45).

Papadopoulos, Theodoros and Yannis Charalabidis (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*. ICEGOV '20. New York, NY, USA: Association for Computing Machinery, pp. 100–111. DOI: 10.1145/3428502.3428514 (cit. on p. 44).

Przeworski, Adam (1991). *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. 1st ed. Cambridge University Press. DOI: 10.1017/CBO9781139172493 (cit. on p. 55).

Radu, Roxana (2021). "Steering the governance of artificial intelligence: national strategies in perspective". In: *Policy and Society* 40.2, pp. 178–193. DOI: 10.1080/14494035.2021.1929728 (cit. on p. 44).
Research, IBM (2018). *Project Debater* (cit. on p. 49).
Risse, Mathias (2022). "Artificial Intelligence and the Past, Present, and Future of Democracy". In: *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Ed. by Oliver Mueller et al. Cambridge Law Handbooks. Cambridge: Cambridge University Press, pp. 85–103. DOI: 10.1017/9781009207898.009 (cit. on p. 44).
— (2023). *Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009255189 (cit. on p. 42).
Romero, Vidal, Beatriz Magaloni, and Alberto Díaz-Cayeros (2016). "Presidential Approval and Public Security in Mexico's War on Crime". In: *Latin American Politics and Society* 58.2. Publisher: Cambridge University Press, pp. 100–123. DOI: 10.1111/j.1548-2456.2016.00312.x (cit. on p. 57).
Salas-Pilco, Sdenka Zobeida (2021). "Comparison of National Artificial Intelligence (AI): Strategic Policies and Priorities". In: *Towards an International Political Economy of Artificial Intelligence*. Ed. by Tugrul Keskin and Ryan David Kiggins. International Political Economy Series. Cham: Springer International Publishing, pp. 195–217. DOI: 10.1007/978-3-030-74420-5_9 (cit. on p. 44).
Schiff, Daniel (2023). "Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy". In: *Review of Policy Research* 40.5, pp. 729–756. DOI: 10.1111/ropr.12535 (cit. on p. 42).
Schiff, Daniel et al. (2021). "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection". In: *IEEE Transactions on Technology and Society* 2.1, pp. 31–42. DOI: 10.1109/TTS.2021.3052127 (cit. on pp. 45, 66).
Schmitt, Lewin (2021). "Mapping global AI governance: a nascent regime in a fragmented landscape". In: *AI and Ethics*. DOI: 10.1007/s43681-021-00083-y (cit. on pp. 42, 44).
Schneier, Bruce, Henry Farrell, and Nathan E. Sanders (2023). "How Artificial Intelligence Can Aid Democracy". In: *Slate* (cit. on p. 49).
Schreier, Margrit (2012). *Qualitative content analysis in practice*. Los Angeles London New Dehli Singapore Washington DC: SAGE (cit. on p. 59).
Smuha, Nathalie A. (2021). "Beyond the individual: governing AI's societal harm". In: *Internet Policy Review* 10.3. DOI: 10.14763/2021.3.1574 (cit. on p. 43).
Sudmann, Andreas, ed. (2019). *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*. 1st ed. Vol. 1. KI-Kritik / AI Critique. Bielefeld, Germany: transcript Verlag. DOI: 10.14361/9783839447192 (cit. on p. 43).

Tallberg, Jonas, Eva Erman, et al. (2023). "The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research". In: *International Studies Review* 25.3. DOI: doi.org/10.1093/isr/viad040 (cit. on p. 45).

Tallberg, Jonas, Magnus Lundgren, and Johannes Geith (2023). *AI Regulation in the European Union: Examining Non-State Actor Preferences*. arXiv:2305.11523 [econ, q-fin]. DOI: 10.48550/arXiv.2305.11523 (cit. on p. 74).

Toll, Daniel et al. (2020). "Values, Benefits, Considerations and Risks of AI in Government: A Study of AI Policies in Sweden". In: *JeDEM - eJournal of eDemocracy and Open Government* 12.1, pp. 40–60. DOI: 10.29379/jedem.v12i1.593 (cit. on p. 74).

Ulnicane, Inga (2022). "Emerging technology for economic competitiveness or societal challenges? Framing purpose in Artificial Intelligence policy". In: *Global Public Policy and Governance* 2.3, pp. 326–345. DOI: 10.1007/s43508-022-00049-8 (cit. on p. 42).

Ulnicane, Inga and Tero Erkkilä (2023). "Politics and policy of Artificial Intelligence". In: *Review of Policy Research* 40.5, pp. 612–625. DOI: 10.1111/ropr.12574 (cit. on p. 69).

Ustundag Soykan, Elif et al. (2022). "A Survey and Guideline on Privacy Enhancing Technologies for Collaborative Machine Learning". In: *IEEE Access* 10. Conference Name: IEEE Access, pp. 97495–97519. DOI: 10.1109/ACCESS.2022.3204037 (cit. on p. 74).

Vaccari, Cristian and Andrew Chadwick (2020). "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News". In: *Social Media + Society* 6.1. Publisher: SAGE Publications Ltd, p. 2056305120903408. DOI: 10.1177/2056305120903408 (cit. on p. 74).

Valle-Cruz, David et al. (2020). "Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation". In: *Government Information Quarterly* 37.4, p. 101509. DOI: 10.1016/j.giq.2020.101509 (cit. on pp. 43, 57).

Verdegem, Pieter (2022). "Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech". In: *AI & SOCIETY*. DOI: 10.1007/s00146-022-01437-8 (cit. on p. 51).

Walker, Christina P., Daniel S. Schiff, and Kaylyn Jackson Schiff (2024). "Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. Issue: 21, pp. 23053–23058 (cit. on p. 50).

Widder, David Gray, Sarah West, and Meredith Whittaker (2023). *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI*. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4543807 (cit. on p. 52).

Wilson, Woodrow (1901). *Democracy and Efficiency* (cit. on p. 57).

Zarkadakēs, Giōrgos (2020). *Cyber republic: reinventing democracy in the age of intelligent machines*. Cambridge, Massachusetts: The MIT Press (cit. on pp. 44, 49).
Zuboff, Shoshana (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. London: Profile books (cit. on p. 51).

2.8 Appendix: List of analysed strategies

Below is the full list of documents that were included in the analysis. The list includes the country, the title of the document, the number of pages, and the publication date. In cases where multiple documents were available that could be considered relevant for this study (e.g., the UK, the US), I opted for the most comprehensive document to ensure that the analysis captures those instances in which governments are most likely to discuss democracy-related issues. The German update in 2020 to its existing 2018 strategy was not considered, as it represents more of a review report. Some countries have published prominent AI policies with a byline such as "report", "whitepaper", "blueprint", or "discussion paper", and not necessarily called these the "national AI strategy". When no clearly designated national AI policy was available, these documents were considered instead (Australia, France, New Zealand, United States). To ensure that these documents can indeed be considered official national strategies, they have been cross-checked against the OECD's global repository of national AI policies and strategies (`https://oecd.ai/en/dashboards/overview`).

- Australia: "Supporting responsible AI: discussion paper" (42 pages), 01/06/2023
- Austria: "Artificial Intelligence Mission Austria 2030" (76 pages), 01/09/2021
- Belgium: "Plan National de convergence pour le développement de l'intelligence artificielle" (58 pages), 28/10/2022
- Chile: "Política Nacional de Inteligencia Artificial y el Plan de Acción" (78 pages), 28/10/2020
- Colombia: "Estrategia Nacional de Inteligencia Artificial" (63 pages), 08/11/2019
- Czechia: "National Artificial Intelligence Strategy of the Czech Republic" (54 pages), 01/05/2019
- Denmark: "National Strategy for Artificial Intelligence" (74 pages), 01/03/2019
- Estonia: "Report of Estonia's AI Taskforce" (47 pages), 01/05/2019
- Finland: "Leading the way into the age of artificial intelligence" (136 pages), 12/06/2019
- France: "For a Meaningful Artificial Intelligence" (154 pages), 28/03/2018
- Germany: "Artificial Intelligence Strategy" (45 pages), 01/11/2018
- Hungary: "Hungary's Artificial Intelligence Strategy 2020-2030" (30 pages), 01/09/2020
- Ireland: "AI - Here for Good" (74 pages), 01/07/2021

- Italy: “Artificial Intelligence at the Service of Citizens” (79 pages), 01/03/2018
- Japan: “Social Principles of Human-Centric AI” (15 pages), 15/02/2019
- Lithuania: “Lithuanian Artificial Intelligence Strategy - A Vision of the Future” (20 pages), 01/03/2019
- Luxembourg: “Artificial Intelligence: a strategic vision for Luxembourg.” (24 pages), 01/05/2019
- Netherlands: “Strategic Action Plan for Artificial Intelligence” (64 pages), 01/10/2019
- New Zealand: “Trustworthy AI in Aotearoa - The AI Principles” (6 pages), 01/03/2020
- Norway: “National Strategy for Artificial Intelligence” (67 pages), 01/01/2020
- Poland: “Policy for the Development of Artificial Intelligence in Poland from 2020” (68 pages), 01/12/2020
- Portugal: “AI Portugal 2030 - An innovation and growth strategy to foster Artificial Intelligence in Portugal in the European context” (40 pages), 01/06/2019
- South Korea: “Policy Tasks for the Stewardship of Trustworthy AI” (33 pages), 11/03/2020
- Spain: “National Strategy for Artificial Intelligence” (89 pages), 01/12/2020
- Sweden: “National approach to artificial intelligence” (12 pages), 01/05/2018
- Switzerland: “Leitlinien ≪Künstliche Intelligenz≫ für den Bund” (11 pages), 25/11/2020
- Türkiye: “National Artificial Intelligence Strategy” (104 pages), 01/08/2021
- United Kingdom: “National AI Strategy” (66 pages), 22/09/2021
- United States: “Blueprint for an AI Bill of Rights” (73 pages), 01/10/2022

Chapter 3

TRANSATLANTIC PERSPECTIVES ON AI POLICY: SHIFTS IN PUBLIC SECTOR PREFERENCES

Abstract: As jurisdictions around the world are drafting policies and regulations for artificial intelligence (AI) technologies, different actors favour different approaches, embracing either innovation-friendly or cautionary positions. To improve our understanding of these different preferences, this paper employs computational text analysis on a comprehensive collection of 317 AI policy documents published between 2016 and 2022 by stakeholders from 56 different countries. It reveals differences and shifts in actors' preferences on AI policy that are structured by geography and sector. Specifically, the analysis shows how the preferences of the public sector in North America and in Europe converged towards each other: while American documents increasingly emphasised topics around protection and harm prevention, the Europeans turned to a more innovation-friendly discourse. These shifts illustrate important changes in governments' approaches towards regulating AI. Notably, the US has moved from a pro-business, laissez-faire approach to the development of more comprehensive – but largely non-binding – policies that reflect an increasingly cautionary stance. In contrast, the EU started off with the goal of developing "trustworthy AI" through a strict regulatory framework. As a consequence, its proposed AI Act has triggered intense discussions around concerns that such far-reaching regulation may hamper the block's competitiveness in AI development. This is reflected in the documents through increasingly innovation-friendly framings in recent years.

3.1 Introduction

Recent advances in artificial intelligence (AI) technology create a strong impetus for policymakers to draft regulations that effectively mitigate risks while maximising benefits. As jurisdictions worldwide grapple with the socioeconomic and ethical implications of AI, an array of stakeholders is engaging in the global policy debate, which informs policymaking and greatly influences the formulation of future AI policies. Within this debate, framing – i.e., "how problems and their potential solutions are articulated and interpreted in policy debates" (Ulnicane 2022) – plays a pivotal role in shaping regulatory action. Consequently, various actors publish documents such as AI ethics guidelines and policy papers to actively engage in deliberative processes of problem definition and agenda setting. It allows businesses, the public sector, and civil society actors to shape perceptions, understandings, and assessments of the technology according to their interests.

Different actors have competing interests when it comes to the regulation of AI (Tallberg, Lundgren, and Geith 2023, Ulnicane et al. 2021). In geographic terms, there are stark differences between jurisdictions (Daly et al. 2019; Hagendorff 2020). Scholarship has identified three prototypical regulatory approaches to AI policy. The US approach is generally business-friendly and prioritises innovation over precautionary restrictions (Roberts, Cowls, Hine, et al. 2021). The Chinese approach has heavily regulated many aspects of AI development but curves out broad exceptions for state-related AI uses (Roberts, Cowls, Morley, et al. 2021). Lastly, the EU is attempting to strike a more balanced third way of "human-centric AI", in which businesses face tougher requirements to ensure precaution, even if it comes at the cost of innovation (Smuha 2019; Roberts, Cowls, Hine, et al. 2021).

In addition to these geographic differences, there are sector-specific preferences. Tallberg, Lundgren, and Geith (2023) have shown that business actors tend to be less concerned about the risks of AI and less supportive of hard regulation than civil society actors. The crucial question that remains open is which positions governments and the public sector assume in all this. Moreover, preferences are not necessarily static. Within the highly dynamic policy debate, actors' perceived and actual interests may evolve over time. Therefore, this paper investigates the following interconnected research questions: how is the global AI policy discourse structured by geography and sector-specific preferences, what shifts in preferences can be identified over recent years, and what does this mean for AI policy outcomes on both sides of the Atlantic?

The empirical analysis contributes to our understanding of the global AI policy debate by painting a nuanced picture of the framings deployed by different actors, and by demonstrating how these map on to the theoretically based differences in preferences across geographies and sectors. Moreover, by including a temporal dimension, it identifies shifts in the discourse that can help explain current policy outcomes and inform political discussions moving forward.

To do so, this paper presents an original data set of such AI policy documents and puts forward a novel sentence-based computational text analysis to investigate different actors' competing interests regarding AI regulation and governance. The document collection covers 317 AI documents from businesses, civil society, and the public sector. Combining dictionary-based methods with zero-shot predictions by a large language model (OpenAI's *GPT-3.5-turbo*), the paper provides sentence-level information on topics and stances (ranging from positive to negative views of AI) for over 50,000 sentences. The empirical analysis allows painting a fine-grained image of the relative prevalence of competing frames, adding to our understanding of the promises and pitfalls different actors associate with the various aspects of AI technology.

Understanding different actors' preferences and framings matters because they play an influential role in the development of new regulations on AI (Tallberg, Lundgren, and Geith 2023, Ulnicane et al. 2021). By shaping the agenda-setting and problem-definition stage of the policy cycle, political actors can condition certain policy outcomes (Elder and Cobb 1984, Gilardi, Shipan, and Wüest 2021).

In the following, I discuss the theoretical and conceptual considerations that inform the research design. I then present the methodology, explaining the data collection and analysis process. The final sections report and discuss the findings.

3.2 Regulatory preferences of different actors

The theoretical starting point of this paper is that different actors have different preferences when it comes to the regulation of AI technologies, and that these preferences are reflected in the way that these actors frame their interventions in the AI policy debate. These interventions can of course take on a variety of forms, but for reasons discussed in the next section, this paper turns to studying AI policy documents published by these actors as an observable manifestation of their preferences. It thus builds on the idea that political actors, when engaging in the early part of the policy cycle, deploy different framings to shape the problem definition and agenda-setting stages (ibid.), which extends to the AI policy debate (Ulnicane et al. 2021, Ulnicane 2022).

This renders AI policy documents an informative source of different actors' preferences regarding AI regulation. In a way, they are "vehicles of messages, communicating or reflecting official intentions, objectives, commitments, proposals, 'thinking', ideology and responses to external events" (Freeman and Maybin 2011). Moreover, studying the prevalent framings deployed by actors matters because these documents are not only an expression of their underlying interests, but through the channels of agenda-setting and problem-definition also have an impact on downstream policy outcomes (Gilardi, Shipan, and Wüest 2021). Ultimately, such documents can reveal information on actors' motivations and practices regarding AI policy (Schiff, Laas, et al. 2022). Accordingly, a number of studies have turned to document analysis of AI ethics principles

(Jobin, Ienca, and Vayena 2019; Fjeld et al. 2020; Hagendorff 2020), government policies (Cath et al. 2018; Radu 2021; Liebig et al. 2022; Djeffal, Siewert, and Wurster 2022) or a mixture thereof (Zeng, Lu, and Huangfu 2018; Floridi and Cowls 2019; Schiff, Biddle, et al. 2020; Daly et al. 2019; Corrêa et al. 2023).

There are many conceivable ways in which different actors' preferences on AI regulation may diverge. One – if not the most – prominent dimension of political conflict is between innovation versus protection, i.e., between more business-friendly and relaxed regulatory approaches on the one hand, and more precautionary, strict regulatory approaches on the other. This conflict may play out between different types of actors (e.g., from the public sector, the business sector, or from civil society), jurisdictions (e.g., the USA, the EU, and China), or a combination of the two axes.

A recent empirical analysis by Tallberg, Lundgren, and Geith (2023) has identified how the innovation vs protection conflict structures policy interventions along sectoral lines. Studying different non-state actors' responses to public consultations on European AI regulation, they find "significant differences across actor types, with business actors being less concerned about the downsides of AI and more in favour of lax regulation than other non-state actors." For the purpose of this paper, I follow their definition of the business sector as including companies, business associations, and corporate-led multi-stakeholder alliances. I label the "other non-state actors" as civil society, which includes non-governmental organisations, social movements, academic institutions, and media outlets. Moreover, I include the public sector into the analysis, which captures all kinds of governmental institutions, from municipal to national level, as well as international (intergovernmental) organisations such as the OECD. This tripartite classification aligns my study with the categories deployed by Schiff, Borenstein, et al. (2021), who in a qualitative review of AI documents detect "meaningful differences across public, private, and NGO sectors". Specifically, they find that public sector and civil society documents cover a wider range of ethical topics and are more engaged with law and regulation than business actor documents.

The specific reasons causing an actor to publish a document on AI ethics or policy are diverse and multifaceted. Schiff, Biddle, et al. (2020) offer a typology of six motivations that might be behind AI ethics documents, though they also acknowledge the difficulty of assessing an actor's true motivation, which might be concealed or overlapping. Indeed, it is almost impossible to properly measure the real motivations and reasons behind a publication process from the outside. However, the purpose of a document and the publishing entity's underlying motives are important considerations that affect the content. To overcome this challenge, this paper relies on the following simplified assumptions in order to interpret the discourse and to draw conclusions regarding what they reveal about different actors' regulatory preferences. First of all, it distinguishes between actors that are clearly interest-driven and those who are more prone to react to those interests. The former includes businesses and civil society organisations, as both of these groups of actors have a relatively straight-forward role to play in public policy

discussions. Broadly speaking, they seek to advance their self-interests (Candler 1999), trying to shape the course of policy by framing issues according to their preferences (John 2021). Publications are one tool that these actors have at their disposal to do so, so the content of these documents can be considered as the product of strategic choices and an editing process that had as its purpose to influence policy making (Schiff 2022). This might also apply to some actors from the public sector, though in general, their starting point is a less clear and predefined set of preferences, especially at the beginning of the policy cycle, which renders them relatively less interest-driven. Thus, these public sector documents – which are generally non-binding in nature – can be considered as expressions of processes that are influenced both by internal considerations and by external pressures, including the actions of business and civil society actors. While this distinction between more and less interest-driven actors may seem minor, it adds important nuance to the interpretation of the subsequent document analysis. It suggests temporal order in which documents from the business and civil society can be understood to have some effect on the content of subsequently published public sector documents. Of course, one can plausibly contest that an opposite mechanism is also at work: businesses and civil society react to agendas set by governments and intergovernmental institutions, and may adapt their subsequent publications accordingly (Shaffer 1995; Hillman, Keim, and Schuler 2004). Still, I argue that their preferences are more clearly structured and therefore less susceptible to variation (see Heckelman and Wilson (2016) on the stability-inducing effect of interest groups on policy in democracies). Therefore, the main observable change should occur in the public sector as a consequence of policymakers leaning more towards one side or the other over the course of time. Another mechanism through which public sector positions may shift is of course a change of government, typically after elections. This sort of exogenous leadership change occurs less frequently for non-state actors.

This analysis is most interested in tracking shifts in the public sector, for two principal reasons. First, as discussed above, this is where we should expect to observe the most interesting variation, which can give hints as to whether businesses or civil society are prevailing in pushing their interests into policymaking. Second, the positions that are prevailing within the public sector are undoubtedly the most informative when it comes to actual policy outcomes, as they usually precede legislative or regulatory proposals, and strongly condition the trajectory of future government policies ().

In addition to sector-specific preferences, competition on the innovation-protection dimension should also be structured along geographic lines. Multiple studies have discussed the different regulatory approaches between the EU, the US, and China (Daly et al. 2019; Cath et al. 2018; Hagendorff 2020; Radu 2021). The focus on these three jurisdictions stems from their elevated role in shaping developments of both the technology and the surrounding political arrangements.

Most observers agree that the three players broadly follow distinct regulatory approaches when it comes to digital technologies and especially regarding AI. Generally,

the EU emphasises a rights- and risk-based ethical framework, which prioritises protection over innovation. It favours binding and comprehensive regulation, and hopes that its regulatory strength will help it set global standards, boost competitiveness, and protect from technological harms, as exemplified by the General Data Protection Regulation (GDPR). This approach is largely reflected in the proposed AI Act and the ongoing political negotiations (Roberts, Cowls, Hine, et al. 2021). In contrast, the US take a more permissive approach which emphasises innovation and self-regulation (ibid.). While it lacks comprehensive federal AI legislation, various states have introduced their AI-related policies, and recently, the Biden administration has displayed more willingness to develop national AI regulation. China, in turn, has rolled out numerous restrictive regulations for AI businesses, while simultaneously prioritising state-centric AI development and deployment for economic and strategic purposes. There is limited emphasis on ethical and individual rights considerations (Roberts, Cowls, Morley, et al. 2021).

Due to methodological constraints, this paper's analysis is primarily concerned with the transatlantic perspective, distinguishing between actors from North America on the one hand, and actors from Europe on the other. While it would undoubtedly be interesting to expand the analysis to actors from other jurisdictions, the lack of sufficient comparable data would put any findings on shaky grounds. Moreover, it seems theoretically implausible to group all these remaining actors together, because their preferences should be quite heterogeneous. For instance, there is little overlap between the Chinese and Japanese approaches to AI regulation, and even less between Chinese and, say, Latin American approaches. However, such a crude regional aggregation would be a necessity for the geographical analysis. Hence, the main reason for focusing on Europe and North America is that for these regions it is more plausible to assume that actors' preferences are somewhat aligned regionally. In addition to documents published by actors that are not from Europe or North America, there are numerous documents in the collection that are published by actors which I consider "global". This could be because they come from multiple authors spanning more than one region, or because the actor is truly supranational in nature (e.g., the UN, the G7, or the OECD). Analysing these documents is undoubtedly relevant, but since it does not fit the specific focus of this paper, they too have been excluded from parts of the analysis.

Putting the main focus on the transatlantic perspective is also pertinent from a public policy angle. When it comes to the regulation of digital technologies, including AI, the EU (and most of Europe, by extension) and the USA are partners and competitors simultaneously. While they may disagree on many details, initiatives such as the EU-US Trade and Technology Council underscore the political willingness to work together in shaping the governance of digital technologies based on shared values and interests. Moreover, the digital ecosystems of both sides are widely integrated through mechanisms such as the Data Privacy Framework and transatlantically operating tech companies. China as the third big player in AI regulation is somewhat isolated from this

	Public sector	Business sector	Civil society
Europe	Precautionary, but more innovation-friendly over time	Innovation-friendly, stable over time	Precautionary, stable over time
North America	Innovation-friendly, but more precautionary over time	Innovation-friendly, stable over time	Precautionary, stable over time

Table 3.1: *Expectations on actors' preferences (baseline and evolution).*

transatlantic area, and its actions are less aligned with the EU and the US. Thus, from a public policy perspective, it can be seen to be playing in a completely different arena, whereas policymakers in America and Europe are more directly competing with their different approaches. This is not to say that the Chinese approach to AI regulation does not matter in the big picture. To the contrary, the country plays a fundamental role in the development of global AI governance. But when it comes to understanding nuances in the Western debate on AI policy, it is most informative to focus on differences and similarities between Europe and America.

One fundamental challenge when comparing documents across different sectors or geographies is that in addition to variation in their preferences, there may be other factors at play that determine the data generation process and thus the final content of their documents. Different sectors and geographies may exhibit structurally different writing styles, and in a reflection of idiosyncrasies in their political systems, they may cater to different audiences and be published with different purposes. The empirical analysis of my data set supports these suspicions. Therefore, the direct comparison across sectors or geographies should be interpreted carefully, and the subsequent analysis is mostly focusing on within-group variation over time, which should control for many of the geographic and sector-specific confounders.

The discussion of geographic and sector-specific preferences allows to draw several expectations regarding the patterns that should be found when analysing AI policy documents. These are summarised in table 3.1 and elaborated on below.

I consider that geography and sector matter independently and in interaction when it comes to shaping an actor's preferences. First of all, there should be overarching patterns concerning sector-specific preferences. By and large, we should expect the business sector to demonstrate markedly more innovation-friendly positions, whereas civil society actors should be focusing on protection (Tallberg, Lundgren, and Geith 2023). For the public sector, it is less evident where it stands. Most likely it will be found in between the other two, though somewhat closer to business preferences given the large influence that economic interests usually have on policymaking (Macher and Mayo 2015). Moreover, when it comes to temporal variation, I expect the public sector to demonstrate a greater degree of variation than the other two, more coherent and interest-driven, blocks. Policymakers working on AI regulation find themselves in a highly contested

environment, drawn to different sides by lobbying efforts from business and civil society actors (Dur, Marshall, and Bernhagen 2019), while bound by public opinion and broader strategic and ideological concerns as well as a global context of competition and geopolitical tension.

Second, regarding geography, it seems plausible to assume that documents in America are generally more tuned to innovation, whereas the debate in Europe is more focused on protection and precaution (see Hammitt et al. (2005) for more on the differences and similarities between US- and European regulatory styles). However, this should be conditional on the sector: in the American context, where the overall debate is highly business-friendly, civil society actors may be even more motivated to discuss ethical issues and thus exhibit more protective positions. In turn, in the European context, where the public sector is already more inclined to take a protective approach, businesses may need to work even harder to emphasise AI's potential and the need for an innovation-friendly environment. These deliberations highlight the importance of looking at variation for each sector and geography separately in addition to overall comparisons.

In addition to expectations about static differences in preferences, this paper takes into account a temporal dimension in its analysis, prompting the formulation of trend expectations. As explained in the next section, the temporal dimension is reduced to a simple binary variable, which assigns documents to one of two periods ("early", from 2016 till mid-2019, and "recent", from mid-2019 till 2022). Regarding businesses and civil society, their preferences should most likely remain relatively stable over time. Even as the public debate evolves, businesses consistently have an interest in a more innovation-friendly regulatory approach, whereas civil society should remain in the opposite camp. However, it seems likely that the public sector will undergo some shifts over the seven year period of this study. The past years have experienced a frenzy of political discussions on how to best address the risks and promises of AI, and the rapid pace of technological advances coupled with the growing salience of AI policy issues have likely caused policymakers in the public sector to adjust their approaches over time. Specifically, it seems plausible to assume that – spurred by a politicisation of AI and changes in the national administration following Biden's election – the public sector in the US has turned slightly more sceptical of unregulated business activities, and over time became somewhat more focused on protection. The Biden administration "seems to be more willing to regulate the free market" (Hine and Floridi 2022), which is exemplified by recent policy developments, such as the release of the Blueprint for an AI Bill of Rights in 2022. In Europe, on the other hand, policymakers had a highly regulation-focused perspective to begin with. This restrictive approach, however, has been subject to intense lobbying efforts by businesses and growing concerns by national governments over limitations to the region's competitiveness in AI development (Justo-Hanani 2022). It is thus plausible to assume that the public sector in Europe has inched towards a more innovation-friendly position over time.

Including the public sector, contrasting European to American perspectives, and detecting shifts in preferences over time all represent important additions to the analysis by Tallberg, Lundgren, and Geith (2023). While this paper builds heavily on the sector-focused approach by Schiff, Borenstein, et al. (2021), it offers novel insights by focusing specifically on the innovation-protection dimension, by disaggregating not only by sectors but also by actors' geographic placement. Moreover, it is one of the first studies in this domain that systematically incorporates a temporal dimension into the analysis. Methodologically, the shift to sentence-based computational text analysis presents a quantitative approach that has so far been underexploited and that can unearth more fine-grained data to complement existing qualitative reviews of AI policy documents.

3.3 Methodology

This section first presents the data collection process which produced the data set on which the paper is based. It also discusses important data processing steps, the computational text analysis techniques, as well as the operationalization of variables that are used in the subsequent analysis.

3.3.1 Data collection

The document collection process builds upon comparable previous research efforts (Fjeld et al. 2020; Jobin, Ienca, and Vayena 2019; Hagendorff 2020). I combined references from various existing data sets and expanded the collection through additional desk research, to capture also more recent documents, covering the period from 2016 to 2022. Following Schiff, Borenstein, et al. (2021), I include a wide range of documents such as frameworks, ethics codes, policy strategies, and reports with policy-relevant sections, as long as they were publicly available and published not by individual authors, but by businesses, public sector institutions, and civil society groups. The large majority of the documents were published in English, though in some instances I turned to translation tools, as described in appendix 3.7.1, which also reports in more detail the document collection process and metadata coding. In terms of focus, many of the documents pertain to the global AI ethics debate, which has largely dominated the first years of AI policy developments (Daly et al. 2019). As with any such collection effort, the inclusion criteria can be expected to significantly alter the subsequent analysis and findings. For instance, the prevalence of business-sector AI ethics frameworks may result in an exaggerated view of how prominent such issues are to the sector. Hence, this analysis can only be an approximation and the generated empirical evidence should always be interpreted carefully, as will also be discussed later on.

The final collection lists 317 documents, which were enriched with additional metadata (e.g., sector, country of publication) based on publicly available information. When-

ever documents were also featured in comparable collections by other researchers, the coding was cross-checked in order to validate it. In the few instances where differences appeared, these were resolved through additional desk research until the most convincing coding was found. Regarding one of the main independent variables of interest for this paper, the sector to which an actor pertains, I opted for a simple categorical classification where a document could only be assigned to one of three options, namely public sector, business sector, and civil society.[1] The public sector captures all kinds of governmental institutions, from municipal to sub-national to national level, as well as international (intergovernmental) organisations such as the OECD. The business sector includes companies, business associations, and corporate-led multi-stakeholder alliances. Lastly, civil society includes non-governmental organisations, social movements, independent academic institutions, and media outlets.

Table 3.2 provides an overview of the 317 documents included in the database and their breakdown by sector and by region. The large majority (163) is issued by public sector entities, whereas business actors and civil society are responsible for 81 and 73 documents, respectively. Most documents in the collection have been published in 2018 and 2019, which can be seen as the prime years of the nascent AI policy discourse.

	Public sector	Business	Civil society	Overall
Europe	71	26	27	124
North America	27	31	24	82
Other	65	24	22	111
Overall	163	81	73	317

Table 3.2: *Count of documents by sector and region.*

I also included information on the country and region of publication, based on the lead authoring organisation's headquarters. Overall, the collection contains documents from 56 countries, though most documents come from just three regions: Europe (124), North America (82) and the Asia-Pacific (64). 31 documents were coded as "global", as they could not be pinned down to one single geography (these include, for instance, communications by the G7 and G20 or by internationally active advocacy groups). Voices from the Global South are clearly underrepresented (only 9 from Africa and the Middle East, and 7 documents from Latin America).[2] As mentioned before, I opted to exclude documents that were not clearly assigned to Europe or North

[1] One could also imagine another operationalization, e.g., a more nuanced scaling system ranging from public to private. But since previous studies generally relied on a simple categorical approach, I chose the same in order to enhance comparability and interoperability.

[2] This unequal representation may in part be due to the researcher's geographic and linguistic constraints (though great efforts have been taken to minimise those during the data collection process). They may also be a reflection of the wider inequalities at play when it comes to ICT capacities and investments in R&D (see Roche, Wall, and Lewis 2023).

America from parts of the subsequent analysis. Including them would have profoundly confused the analysis due to the small number of cases falling into the respective categories, and due to the suspected lack of coherent preferences.[3] Table 3.3 provides an overview of the restricted sample of 206 transatlantic documents, and also shows how many documents of each subgroup fall into which period. To code the binary temporal variable, I took the median publication date (2019/04/17) as a cut-off point, as it coincides with a pivotal moment in the development of global AI policy. Notably, this moment is characterised by the publication of several landmark documents such as the US government's American AI Initiative in February 2019 or the European Commission's AI High-Level Expert Group's Principles in April 2019, followed by the OECD Principles the next month (see Baum et al. 2023). Then, documents published on or before that date are considered to be from the "early" period, whereas those published afterwards are "recent".

	Public sector	Business	Civil society
Europe	33 / 38	14 / 12	12 / 15
North America	13 / 14	14 / 17	17 / 7
Overall	46 / 52	28 / 29	29 / 22

Table 3.3: *Count of documents contained in the restricted sample, by sector, region, and period (earlier / more recent).*

3.3.2 Computational text analysis

To reveal actors' preferences as expressed through documents, I turn to computational text analysis, notably dictionary-based methods and zero-shot predictions by a large language model (LLM). Information is extracted at the sentence level, before being aggregated to document-level variables on topic attention and stance.[4]

The information on topic attention is used to compute a variable I call *innovation affinity*. It measures the degree to which a document speaks about innovation-related topics relative to protection-related topics.[5] The variable ranges from 0 to 1. A higher value implies more attention to innovation topics, which I take as a proxy for business-friendly positions. This is of course no perfect measure, as it is plausible to imagine

[3] Where I do list them (e.g., in 3.2 and in the appendix), this is aimed to provide an additional benchmark which helps situating the European and American numbers into a more global context.

[4] While stance is related to more commonly known sentiment analyses, Bestvater and Monroe (2023) have shown that there are important differences, especially in political contexts. For studies that are interested in the underlying political attitudes rather than the overall tone of writing, it is therefore crucial to assess stance with respect to a given target. In this case, the targeted object would be "AI technologies".

[5] To compute *innovation affinity*, I divide a document's share of sentences speaking about innovation by the sum of the shares of sentences speaking about innovation and/or protection: $affinity_{inno} = share_{innovation}/(share_{innovation} + share_{protection})$

other motivations that could drive topic attention. For instance, some civil society actors might speak heavily about innovation in order to convey how it may harm society. Or a business actor may choose to focus excessively on ethics as a means of virtue signalling or so-called AI ethics-washing. With this limitation in mind, I nevertheless consider that the resulting measure gives a good indication of an actor's preferences. In the appendix, I report several validity checks that I conducted. One advantage of using relative attention as a measurement is that it controls for document-specific characteristics that might obstruct other measures. For instance, the documents vary greatly in length, from a couple of sentences to hundreds of pages (see table 3.6 in appendix 3.7.1). Since both parts of the attention ratio are relative shares, they cancel out the effect of document length on the measurement.

Stance refers to the position of an actor towards AI, as expressed in a sentence. Concretely, a sentence may reflect a more positive, neutral, or negative view of the technology. Having such information not only allows quantifying the overall stance expressed in a document by a stakeholder but, by crossing it with the topic information, also reveals crucial topic-specific variation.

To enable computational text analysis of the documents, the texts were converted into a machine-readable format. Moreover, non-informative elements had to be stripped (such as page numbers, footnotes, or masthead/boilerplate information). In cases where the text was only available as a non-readable PDF or some other file format, text was manually copied and re-formatted to ensure comparable data quality across documents. For very lengthy documents (> 50 pages), I restricted the text to executive summaries (where available) or a selection of key chapters/policy-relevant sections. As I analyse relative shares rather than absolute counts, I do not expect this choice to be problematic.

The documents were then tokenised into sentences via R, resulting in 53,029 units. For the dictionary-based approach, individual sentences have been pre-processed following standard methods (uncasing, removal of punctuation, numbers, and special characters). Then, to detect whether a sentence spoke about a given topic or not, I used a dictionary lookup approach. Basically, it labels a sentence as speaking about innovation when a related keyword (such as "investment", "startup", "sandbox" or "intellectual property") is detected. For the protection topic, I opted for a slightly more granular approach – using several dictionaries to first code for sub-topics before collapsing the information into an overall protection variable. I did this to account for the topic's more complex and multifaceted nature and to allow for a more granular future analysis. For the ethical sub-topics, I took the themes identified by Fjeld et al. (2020), such as transparency, accountability, or safety. Notably, sentences may address both innovation and protection simultaneously. In such cases, both labels were assigned to sentences.

Appendix 3.7.2 explains the procedure in more detail, listing the topic dictionaries together with benchmarking results comparing the predictions to a validation sample labelled by human coders in order to ensure output validity. Overall, the accuracy of

the dictionary-based classification tasks ranges from 82% to 98%, and is consistently on par with human intercoder reliability.[6]

Dictionary-based approaches can work well for identifying topics, but they tend to perform quite poorly for sentiment or stance detection, especially in political documents (Muddiman, McGregor, and Stroud 2019; Atteveldt, Velden, and Boukes 2021). These texts typically lack the strongly opinionated signals that we may find, for example, in tweets or online reviews. Instead, institutional authors express their views and opinions in more subtle, nuanced, and indirect ways. Often, there are large editing teams involved in the production of such texts, who have to negotiate input from a range of different internal contributors. To overcome the limitations of dictionary-based approaches, I again turned to fine-tuning a more sophisticated BERT classifier. However, the quality of the output was unsatisfying, so I decided to deploy a novel approach based on an LLM's zero-shot learning capabilities. Kheiri and Karimi (2023) report very high accuracy when prompting GPT-3.5-turbo to classify tweets according to sentiment. And Gilardi, Alizadeh, and Kubli (2023) have recently demonstrated that LLMs such as ChatGPT can outperform human coders for routine annotation tasks such as stance or frame detection. Following their approach, I tasked OpenAI's *GPT-3.5-turbo* model with classifying each sentence according to whether it expressed positive, neutral, or negative views towards AI, scoring its responses on a 5-point Likert scale. To minimise potential issues of randomness or hallucination, I repeated the process three times and took the average score as the final prediction.[7]

Such AI-generated output should be dealt with carefully by researchers. Appendix 3.7.3 provides a detailed explanation of the procedure, the exact prompt, and measures of output reliability as well as additional output validation based on human review. Overall, while not perfect, I consider the LLM output to be sufficiently accurate and reliable to produce sentence-level estimates of actors' expressed stances towards AI. The model performs well in most cases, and the instances in which it errs are systematically biased towards assessing sentences as slightly more positive than a human reviewer would. As this slight positive bias should apply systematically to all sentences, it should not distort the region- or sector-specific comparisons too much. Moreover, I am mainly interested in relative levels of stance (e.g., across sectors and regions, or over time) for the downstream analysis, which further mitigates the bias. Therefore, the next sections report the results of the topic and stance analysis.[8] I then discuss the findings' implications for our understanding of how different actors frame AI policy.

[6]Since estimations for very short documents are more sensitive to individual classification errors, I manually reviewed those, as explained in appendix 3.7.6.

[7]I also experimented with an analogous ChatGPT pipeline for the topic detection task outlined above, but the ambiguous and domain-specific criteria of the sub-topics resulted in mediocre performance, so I opted for the more transparent and reproducible dictionary approach.

[8]To illustrate the outcome of the automated text analysis, an extract with example sentences together with their corresponding labels and scores can be found in appendix 3.7.4.

3.4 Findings

3.4.1 Shifts in attention

Table 3.4 reports the (static) distribution of topics across sentences, breaking it down by sectors and regions.[9] Across all documents, the protection topic was relatively more prevalent (detected in 23.1% of sentences) than the innovation topic (9.2% of sentences). As many of the documents in the collection strongly focus on AI ethics, this should not be surprising. But it also suggests that the global AI policy discourse is quite concerned with topics of protection (at least on paper).

	(1) Innovation	(2) Protection	(3) Both	(4) None	(5) Innovation affinity
Overall	9.2	23.1	1.5	69.2	28.5
Public sector	11.3	19.4	1.6	71.0	36.8
Business sector	4.4	31.6	1.3	65.3	12.2
Civil society	4.5	31.6	1.3	65.1	12.5
Europe	10.7	20.1	1.5	70.8	34.7
North America	7.2	29.6	1.9	65.1	19.6
Others	8.4	23.6	1.3	69.3	26.2

Table 3.4: *Share of sentences (in %) speaking to innovation, protection, both, or none, as well as innovation affinity. Average scores by sector and region.*

The table shows that variation regarding the share of sentences mentioning innovation (column 1) is larger across sectors than across regions, with the public sector devoting much more attention to it. Notably, both civil society and businesses have a very low share of attention to innovation, which in the case of the former goes against a priori expectations. This is most likely an artefact of the type of document published by these actors, which often are ethics principles naturally more geared towards the protection topic. A similar logic can help understand why attention to the protection topic (column 2) is very high for businesses, as well as civil society.

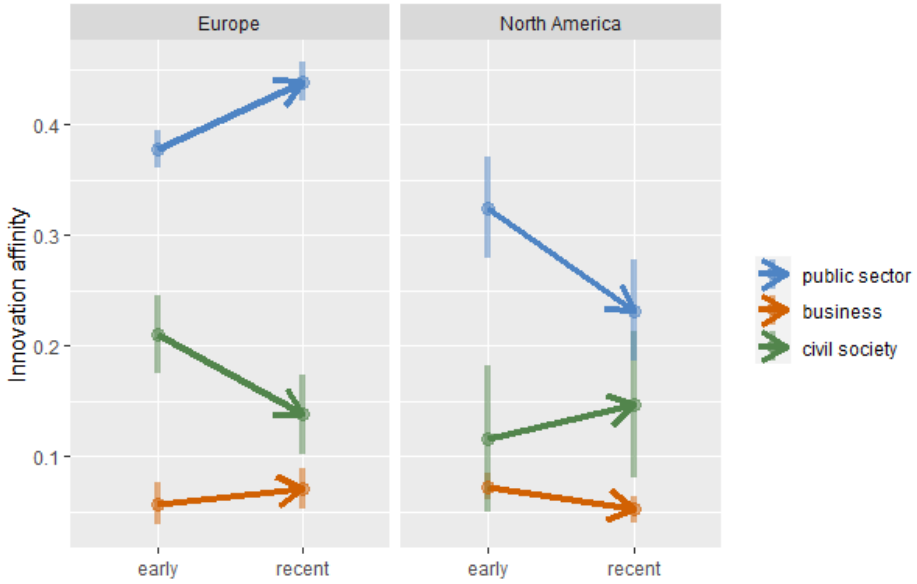
It stands out that the protection topic experiences high variation both across sectors and across regions, with documents published by actors from North America exhibiting a much higher focus on this than other regions. Regarding the share of sentences that address both protection and innovation (column 3) or neither (column 4), estimates are relatively consistent across sectors and regions.

Column 5 in table 3.4 reports documents' mean innovation affinity. As explained in the previous section, it is computed as a document's share of sentences speaking about innovation, relative to all "topical" sentences (i.e., those speaking to innovation and/or protection). A value of 0 implies that a document's topical sentences have exclusively

[9] To account for differences in document length which may distort the results, I repeated the calculations by first aggregating at the document level. While point estimates change, the overall pattern remains robust, as reported in the appendix 3.7.5

covered protection, whereas a value of 1 (i.e., 100%) implies that they have exclusively covered innovation. We can see that the innovation affinity is much lower for both business and civil society actors, implying that they are relatively more concerned about protection than innovation. Again, this seems somewhat counter-intuitive at first sight regarding the business sector, whose preferences were expected to be more tuned towards innovation. The following section will discuss this in more detail and offer some explanations.

Figure 3.1: *Mean attention to innovation versus protection, by sector, region, and over time.*



Note: Innovation affinity reports the average share of documents' sentences on innovation versus sentences on innovation and/or protection. A value of 0 implies that relevant sentences in a document have exclusively covered protection, whereas a value of 1 implies that they have exclusively covered innovation. Error bars indicate the 95% confidence intervals for differences between two time periods.

To visualise temporal developments, figure 3.1 plots the evolution of average attention to innovation versus protection for the three different actor groups, over the two periods, and split by region.[10] First of all, we can see that of the three sectors, the public sector displays by far the most concern with innovation relative to protection. This holds for both regions, as well as the wider sample covering documents from other regions (see appendix 3.7.5). It can probably be explained to a considerable degree by the different composition of documents published by each sector. Whereas a large number

[10]Notably, the observed patterns are robust also across different regional group specifications (e.g., restricting Europe to only EU member states and North America to only US-based actors).

of documents by businesses and civil society are AI ethics principles predominantly focused on topics related to the protection category, the public sector comprises numerous comprehensive policy strategies, which almost by definition dedicate more attention to investment, R&D, and other innovation-related topics. This serves to illustrate the limitations in comparing attention shares directly *between* sectors, and justifies the focus on temporal change *within* a given sector.

In that regard, figure 3.1 shows that, generally in line with my expectations, the public sector exposes the most adjustment in how much attention is paid to innovation relative to protection. A similarly sized adjustment can be found only for civil society in Europe. Notably, the shifts in the public sector when comparing Europe to America are divergent, pointing in different directions. They are also divergent when comparing civil society with the public sector within each region, whereas businesses point (very slightly) in the direction of the public sector. Whereas in Europe, civil society recently become more attuned to protection, the public sector has clearly shifted towards innovation. The opposite is the case in North America, where the attention to innovation displayed by public sector documents has experienced a remarkable drop. At the same time, civil society actors have become somewhat more alert to innovation. The business sector is remarkably constant in both regions, demonstrating the lowest share of attention to innovation of all actors. As mentioned before, this should be predominantly due to it covering mostly lofty AI ethics principles.

The observed changes in the public sector are quite substantive. In Europe, we record a .065 increase on the normalised $0-1$ attention scale. Whereas in earlier documents, out of 100 topical sentences (i.e., sentences on innovation and/or protection), 38 were on innovation (and 62 on protection), more recently this jumped to 44, implying a 16% increase. In the case of American public sector documents, the .1 drop implies a 30% decrease. Remarkably, this shift by the public sector is larger than the overall difference between civil society and business sectors in either period. However, not all of the observed changes are statistically significant. Given the small sample sizes, this is hardly surprising. The error bars around the dots indicate the 95% confidence intervals. No statistical significance can be determined for the shifts of civil society in North America, nor for businesses in Europe. This reminds us of the exploratory nature of this research, whose findings should be seen as indicative, but not conclusive.

Innovation affinity is a relative measure that captures the share of sentences focused on innovation (in the numerator) versus the share of sentences focused on protection (in the denominator). Hence, a change could be caused by underlying changes in either (or both) components of the fraction. In order to get a better understanding of what is driving these dynamics, I also checked for changes in both components separately. It turns out that for the public sector, in both Europe and America, the overall change is driven by the simultaneous, but divergent, changes in attention to protection and innovation. This is different to civil society and businesses, where attention to both innovation and protection has been growing, but to different degrees. This gives some

further support to the notion that the public sector has undergone the strongest adjustment in its preferences, whereas the observed changes for the interest-led actors are mostly driven by changes to their overall document structure (more sensitive to both dimensions of interest overall).

The protection topic combines several sub-topics (taken from Fjeld et al. 2020), as explained previously. To zoom in and see whether there is variation over time in the degree to which different sub-topics are covered, figure 3.2 plots the share of protection-related sentences dedicated to this sub-topic (using the full sample of documents).[11] While fully disentangling these would go beyond the scope of this paper, it seems worthwhile to highlight some trends within sub-topics that stand out. First of all, the theme of *fairness and non-discrimination* has remained remarkably stable and highly prominent throughout the years. *Accountability* and *safety & security* follow closely, though with a bit more variation. The prominent role of these three themes reflects earlier findings by Jobin, Ienca, and Vayena (2019), Fjeld et al. (2020) and others. *Transparency & explainability* are a bit less prevalent, and share with *accountability* and *safety & security* the slight downturn in recent years – though this may be more noise than signal.

The data gives more robust evidence for two final observations, which relate to the rise of attention to *privacy* and *human control over technology*. The latter was broadly absent from earlier documents but has received increased focus in recent years (though still at relatively low levels). Notably, this uptick is observed most strongly among businesses and civil society, but much weaker within the public sector. Even more drastic is the increase in attention to *privacy*, which is surprising given that Jobin, Ienca, and Vayena (2019), Fjeld et al. (2020) already find it addressed in many landmark documents. It may be that earlier documents have acknowledged the issue, e.g. through references to data protection guidelines, but have not discussed it in depth.

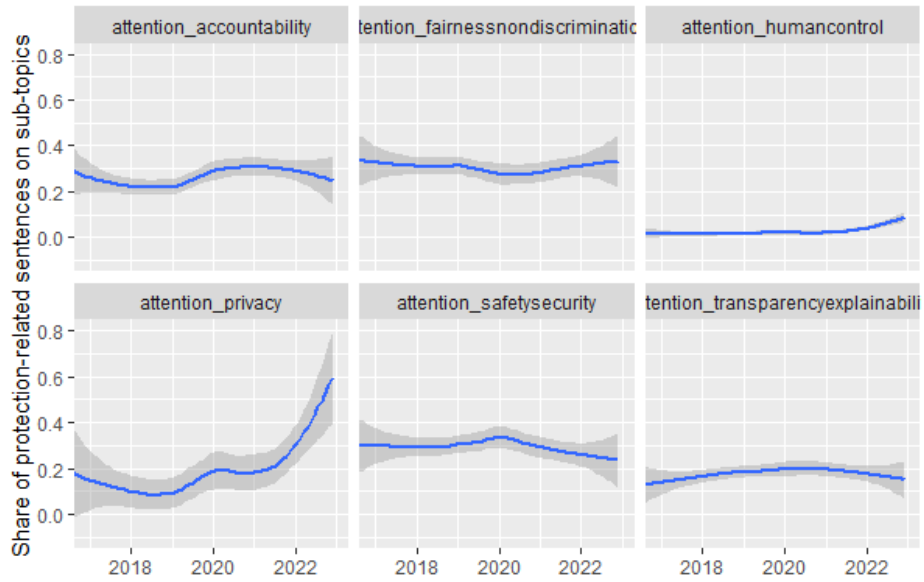
	Public sector	**Business sector**	**Civil society**
Europe	More innovation-friendly over time	Stable over time	More precautionary over time
North America	More precautionary over time	Stable over time	Stable over time

Table 3.5: *Observed evolution of actors' preferences on AI regulation.*

Table 3.5 summarises the main takeaways that can – cautiously – be drawn from the analysis. There is evidence supporting some, but not all, of the principal expectations underpinning this paper: first, there seem to be relevant differences between sectors and regions, which should be studied carefully. Second, when it comes to overall levels of attention, the findings do not always map onto the expectations. It is unclear whether this contradicts the expectations or whether it might be a result of the data collection and

[11] Disaggregating further by region or actor results in very fluctuating graphs that are hard to meaningfully interpret.

Figure 3.2: *Share of protection-related sentences dealing with individual sub-topics.*



analysis process, which highlights the methodological limitations. Third – and in line with expectations – the public sector in Europe has grown more innovation-friendly over time, while the opposite has occurred in America. This can be seen in the shifts in attention to innovation- versus protection-related topics. Fourth, while preferences of civil society and business actors are not fully static either, they demonstrate more stable estimates over time. Ultimately, this fits in with the assumed distinction between the more interest-driven business and civil society sectors on the one hand, and the less interest-driven public sector on the other hand.

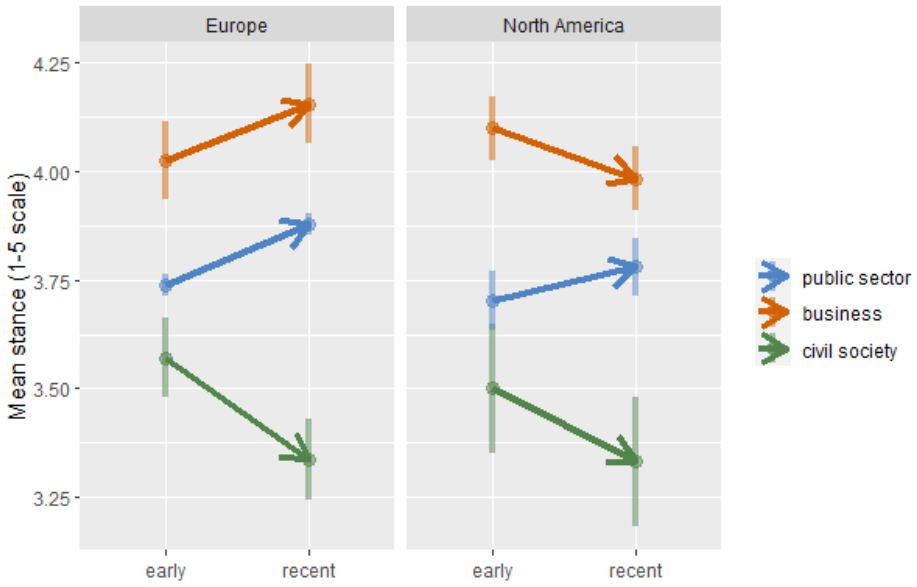
3.4.2 Shifts in stance

To provide additional insights into the shifts in actors' preferences regarding AI regulation, the documents were also analysed regarding the stance they express regarding the technology. The idea is that a more negative stance is reflective of concerns and thereby indicative of an actor leaning more towards protection. In turn, a more positive stance reflects the aspirations associated with the technology and suggests that the publishing institution leans towards an innovation-friendly approach.

Figure 3.3 plots the evolution of the stance with which AI is covered in the documents, again broken down by actor and region.[12] It reveals substantial differences between sectors, some of which are moderated by a region's temporal trends. Overall, it

[12] Appendix 3.7.5 extends this figure beyond transatlantic actors.

Figure 3.3: *Mean stance across sentences, by sector, region, and over time.*



Note: Stance based on averaged GPT3.5-turbo predictions. Error bars indicate the 95% confidence intervals for differences between two time periods.

can be seen that businesses act as AI's cheerleaders whereas civil society serves a critical watchdog function. In that, the pattern seems to reflect actors' overall preferences for less and more regulation, respectively. The public sector ranges somewhere in between the other two, though more recently, public sector documents in Europe have even overtaken businesses in positive framing, suggesting an embrace of the technology by policymakers. However, static comparisons between sectors should be interpreted carefully, as different levels might also be the result of different overall writing styles that are unrelated to an actor's specific stance on AI. Presumably, publications by businesses are written more positively than civil society organisations in general, regardless of the topic. Thus, the focus of the subsequent analysis is again on comparing developments *within* rather than *across* sectors.

In that, figure 3.3 shows the remarkable shift of the public sector in Europe, which has become much more positive when speaking about AI, mirroring its increased innovation affinity. No such statistically significant increase has been observed for the American public sector. In both Europe and North America, civil society has moved towards even more negative stances. However, this shift is only significant in Europe. An interesting pattern can be observed for the American business sector, which has shifted towards a more critical tone in recent years. In Europe, businesses are paralleling trends in the public sector, albeit at more positive levels.

In general, the recorded shifts are more subtle than in the topic attention figures plotted above. But still, some of them are sizeable and statistically significant. In Europe, the public sector has more recently employed language that is on average 0.12 points more positive, as rated on a 5-point scale. In other words, sentences are about 3% more positive in recent documents. Considering the nuanced nature of the language typically employed in such documents, even such a small change is telling. The other statistically significant change is the drop in stance for European civil society documents, which went from 3.57 to 3.34. The shifts recorded for North America are more marginal and not statistically significant.

Overall, the stance analysis reflects most, but not all, of the expectations and earlier findings. Compared to the takeaways described earlier in table 3.5, the more precautionary stance of the American public sector does not come with a more negative framing. Also, both civil society and businesses exhibit greater variation in the stance analysis, though the changes are bordering on statistical significance and might also be noise. Where the stance analysis does speak very clearly to the expectations is regarding the overall levels of support for/opposition to AI, as expressed through the business sector's more positive and civil society's more negative framing.

3.5 Discussion

Evidently, both approaches – attention and stance analysis – are just approximations, highlighting the need for careful triangulation when studying such document collections through computational text analysis. Still, they allow for several insights regarding trends and patterns, which can inform discussions on transatlantic AI policy development. The following section initiates this discussion by contextualising the findings and by offering additional observations and interpretations.

Taken together, the findings broadly reflect current developments in AI policy on both sides of the Atlantic, as well as the translatlantic dynamic of two competing, but somewhat convergent approaches to AI regulation. During negotiations for the EU's AI Act, the importance of remaining open to innovation has been stressed by numerous stakeholders. National governments such as Germany and France are eager to ensure that their domestic technology companies maintain competitive vis-à-vis Chinese and American competitors. This has been exemplified by the difficult negotiations on whether the AI Act should also target so-called foundational models, such as ChatGPT. For the same reasons, business representatives are keen on minimising regulatory burdens, and through massive lobbying have aimed to push the debate towards more innovation-friendly positions. These developments can be seen in the content of public sector documents, and they are also hinted at by the upwards shift in the business sector. Notably, though, despite these slight upward shifts towards more focus on innovation, the protection topic remains the most-addressed topic across documents, suggesting

that AI ethics and cautionary considerations retain a prominent spot in the discussion.

In contrast to Europe, the US government was initially very open to innovation and sceptical of any regulation that would slow down the development of AI.[13] In recent years, and likely pushed by Joe Biden assuming the presidency in 2021, documents published by the public sector in North America have demonstrated a higher willingness to engage with protection-related topics. In 2022, the White House published a "Blueprint for an AI Bill of Rights", which stood in stark opposition to the laissez-faire approach which was dominant during the Trump administration. The White House's recent Executive Order "Safe, Secure and Trustworthy Development and Use of Artificial Intelligence" is the most notable outcome of this shift, as it displays a much higher commitment to protecting Americans from AI, even if that comes with some regulatory burdens that might slow down innovation.

While there are clear and prevailing differences between the American and European approaches to regulating AI, recent developments indicate a certain alignment or convergence.[14] In addition to the above-mentioned domestic shifts, convergence can also be detected in bilateral relations. The structured dialogue through the EU-US Trade and Technology Council serves as a case in point. Established in June 2021, it aims to promote, inter alia, transatlantic cooperation on AI. Notably, as a result of this cooperation, the two sides agreed on a "Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management" in December 2022, with the expressed objective of "[b]ringing EU and U.S. approaches closer" (European Commission, 2022). The Joint Statement following the fourth ministerial meeting in May 2023 reiterated this high-level commitment to close cooperation and coordination. Greater ideological alignment between the Biden administration and European policymakers is one likely factor to explain this convergence. Another is the transatlantic alliance's desire to limit China's influence in the field of AI development and governance. There are signs that Beijing is reacting to the closer transatlantic cooperation and alignment. To the surprise of many observers, during the AI Safety Summit the Chinese government has signalled its willingness for more international dialogue and also recently gave in to European demands on cross-border data transfer rules.

Beyond the public sector, there are some noteworthy observations for the business sector that are somewhat at odds with the conventional wisdom that these actors' are more focused on innovation. In contrast, it displays a low innovation-affinity (and, by extension, a higher relative focus on protection). This might speak to the strategic considerations underpinning the publication process. For instance, actors may conceal

[13] Notably, there are occasional policy initiatives at the municipal or state level which demonstrate a more restrictive approach to AI regulation, but in the big picture, the US did clearly have a more permissive approach.

[14] This is not immediately visible from the raw data, which actually suggests a divergence of viewpoints between the two regions' public sectors. Instead, it emerges from contextualising the observed shifts by taking into account broader political developments.

their true preferences and engage in strategies such as AI ethics-washing, emphasising protection-related themes as a way of virtue-signalling to external and internal stakeholders (Attard-Frost, De los Ríos, and Walters 2022). By signalling commitment to ethical principles (on paper), corporate actors downplay the urgency for regulatory interventions. Indeed, non-binding self-regulation and industry-led technical standards have been the predominant steering tools over the last years. Only more recently have sharp legislative developments such as the EU's AI Act taken centre stage. At the same time as it speaks a lot about protection, the business sector also puts forward the most positive framing of AI technology overall. This seems plausible given that it is mostly composed of companies or corporate actors developing and/or using AI. The positive framing may serve multiple purposes, such as signalling opportunities to investors and business partners or securing support from policymakers and the general public.

In this context, the global AI debate has recently paid a lot of attention to issues around *human control of technology*. Curiously, though, this is the least covered of the different protection-related ethical themes for the period under study (see table 3.7 in appendix 3.7.2). While this is broadly in line with other work that has identified this to be a lower-priority theme for AI documents (see, for instance, Fjeld et al. 2020), it is a stark mismatch from the state of global AI policy discussions in 2023, in which the question of superhuman or general AI – so powerful that it will cause humans to lose control of the technology or otherwise experience catastrophic risks – features prominently. A high-level example of this is the AI Safety Summit, an international conference convened by the UK government in early November 2023 and pitched to discuss safety and regulation of so-called "frontier AI". The summit explicitly aimed to focus on the misuse of AI and the risks of loss of control.

On the one hand, this latest shift in attention towards issues of human control may be explained by the most recent breakthroughs in generative AI tools, e.g., large language models such as ChatGPT and image creation tools such as Midjourney. As these have become more powerful and much easier to access, people are starting to get concerned about technology getting out of control. Since this analysis is restricted to documents from 2016 to 2022, it captures a moment of the discussion in which these advances still seemed relatively abstract and farther away. On the other hand, the recent shift of attention towards themes around human control may be a strategic choice by business actors. It is a recurring argument in AI ethics debates that the focus on long-term dangers deriving very powerful superhuman or general AI distracts from AI's immediate, tangible dangers. By pushing such a framing, business actors may hope to distract lawmakers and the broader public from the very real short-term consequences of their AI tools, thus avoiding or downplaying impending regulation. Hence, prominent NGOs and observers have repeatedly blamed business representatives and AI ethicists that are close to what is often referred to as existential or catastrophic risks camp for sounding the alarm about some hypothetical, relatively far-away future in order to un-

dermine efforts for concrete and rapid policy interventions.[15] At the same time, this paper's document analysis suggests a similar trajectory for civil society as for businesses when it comes to emphasising these questions.

While the analysis allows for several important insights that can inform our understanding of different actors' regulatory preferences on AI, it is not without limitations. These imply caution when interpreting the findings, which should be understood as indicative and exploratory, rather than conclusive. For one, the limited size of the document collection results in relatively small subgroups when disaggregating for sector, region, and time period. Accordingly, many of the observed changes over time are not statistically significant, but merely give us a sense of the direction in which developments are probably pointing. Moreover, the analysis gives the same weight to each document, which is hardly reflective of the variegated importance that authors and audiences attribute to individual texts, or the impact a text has due to its legal nature. For instance, a single hard law passed to foster AI innovation may effectively outweigh dozens of noble AI ethics declarations when it comes to actual policy outcomes. Hence, a more granular analysis might try to control for this by operationalizing measures for "importance" of documents, though this would arguably overstretch resources and be unattainable for this sort of empirical data.

Importantly, already now the effect sizes, especially for stance, are quite marginal. While I argue that – given the technical and political nature of these carefully drafted documents – even small changes are an expression of broader underlying shifts, the findings should not be overinterpreted. More technical limitations relating to the document collection and computational text analysis methods are discussed in appendix 3.7.6.

Keeping these limitations in mind, the analysis presented in this paper nevertheless offers an important methodological and substantive contribution. It puts forward a novel, sentence-based content analysis of AI policy documents, revealing shifts in actors' attention and stance, which can allow insights into their underlying preferences. As the discussion of the findings has shown, the shifts in the public sectors in Europe and North America can be seen as indicative of the downstream policy outcomes we are currently observing.

In sum, the results provide valuable exploratory insights that are indicative of different actors' preferences for AI regulation and the evolution of their positions over time. To enrich our understanding of the global AI policy debate, future research could build on the insights generated by this paper, expanding the scope of the analysis to actors beyond Europe and North America, and enriching it further by covering additional documents or other text data sources. This paper has shown that the focus on the innovation-protection dimension as one of the critical lines of conflict in AI regula-

[15] See, for instance, a collection of statements by leading AI experts published by the Center for AI Safety (`https://www.safe.ai/statement-on-ai-risk`, accessed on 12/06/2023).

tion can be fruitful beyond existing applications (Tallberg, Lundgren, and Geith 2023). Moreover, there are certainly other conflicts which computational text analysis methods might help to identify and investigate.

3.6 Conclusion

As AI technologies advance at an accelerating pace, policymakers around the world are developing AI regulations. The accompanying policy discussions are informed, inter alia, by AI ethics and policy documents published by stakeholders from the public and business sectors, and civil society. To learn more about these actors' preferences, and especially about critical differences between the American and European approaches to regulating AI, this paper puts forward novel computational text analysis methods deployed on a large document collection of 317 relevant documents, spanning from 2016 to 2022.

The analysis reveals several significant differences between sectors and regions when it comes to the attention they pay to innovation versus protection – a critical conflict regarding AI regulation, as identified by previous research (ibid.). Moreover, there are notable differences when it comes to the stance that documents convey on AI technologies, i.e., whether they are framed in a more positive or negative way. One of the paper's key empirical contributions is that it allows tracing shifts in attention and stance over time, thus providing a more nuanced and fine-grained understanding of the dynamic AI policy debate. In line with expectations, the public sector in Europe has moved towards a more innovation-friendly position over time, whereas the American public sector has become more concerned with protection. As expected, the business and civil society sectors, which can be seen as more interest-driven, are more static in their revealed preferences, with the exception of civil society in Europe, which has become ever more focused on protection and more negative in its stance.

Taken together, these findings reflect current developments in AI policy on both sides of the Atlantic. During negotiations for the EU's AI Act, the debate in Europe has become more susceptible to arguments about innovation, as businesses and governments feared that an overly restrictive regulatory approach might hinder the block's competitiveness and innovation capacity. In the US, conversely, the new administration under Joe Biden has shown a larger appetite to regulate technology firms, even if it comes at the expense of restraining innovation. This underlines a certain convergence by the two blocks, with promises that transatlantic cooperation on the regulation of digital technologies, and especially on AI, may become more aligned. Considering the current geopolitical climate, which pits the transatlantic partnership against an increasingly assertive China, such regulatory alignment for what is arguably going to be one of the next decades' most transformative technologies appears sensible.

Bibliography

Attard-Frost, Blair, Andrés De los Ríos, and Deneille R. Walters (2022). "The ethics of AI business practices: a review of 47 AI ethics guidelines". In: *AI and Ethics*. DOI: 10.1007/s43681-022-00156-6 (cit. on p. 107).

Atteveldt, Wouter van, Mariken A. C. G. van der Velden, and Mark Boukes (2021). "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms". In: *Communication Methods and Measures* 15.2, pp. 121–140. DOI: 10.1080/19312458.2020.1869198 (cit. on p. 98).

Baum, Kevin et al. (2023). "From fear to action: AI governance and opportunities for all". In: *Frontiers in Computer Science* 5, p. 1210421. DOI: 10.3389/fcomp.2023.1210421 (cit. on p. 96).

Bestvater, Samuel E. and Burt L. Monroe (2023). "Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis". In: *Political Analysis* 31.2. Publisher: Cambridge University Press, pp. 235–256. DOI: 10.1017/pan.2022.10 (cit. on p. 96).

Candler, Gaylord George (1999). "Interest Groups and Social Movements: Self- or Public Interested? Insights from the Brazilian Third-Sector Literature". In: *Voluntas: International Journal of Voluntary and Nonprofit Organizations* 10.3, pp. 237–253. DOI: 10.1023/A:1021205017395 (cit. on p. 90).

Cath, Corinne et al. (2018). "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach". In: *Science and Engineering Ethics* 24.2, pp. 505–528. DOI: 10.1007/s11948-017-9901-7 (cit. on pp. 89, 90).

Corrêa, Nicholas Kluge et al. (2023). "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance". In: *Patterns* 4.10. DOI: 10.1016/j.patter.2023.100857 (cit. on p. 89).

Daly, Angela et al. (2019). *Artificial Intelligence, Governance and Ethics: Global Perspectives*. SSRN Scholarly Paper ID 3414805. ZSCC: 0000000. Rochester, NY: Social Science Research Network (cit. on pp. 87, 89, 90, 94).

Djeffal, Christian, Markus B. Siewert, and Stefan Wurster (2022). "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strate-

gies". In: *Journal of European Public Policy* 29.11, pp. 1799–1821. DOI: 10.1080/13501763.2022.2094987 (cit. on p. 89).

Dur, Andreas, David Marshall, and Patrick Bernhagen (2019). *The Political Influence of Business in the European Union*. Ann Arbor, MI: University of Michigan Press (cit. on p. 93).

Elder, Charles D. and Roger W. Cobb (1984). "Agenda-Building and the Politics of Aging". In: *Policy Studies Journal* 13.1, pp. 115–129. DOI: 10.1111/j.1541-0072.1984.tb01704.x (cit. on p. 88).

Fjeld, Jessica et al. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society (cit. on pp. 89, 94, 97, 102, 107, 114–116).

Floridi, Luciano and Josh Cowls (2019). "A Unified Framework of Five Principles for AI in Society". In: *Harvard Data Science Review*. DOI: 10.1162/99608f92.8cd550d1 (cit. on p. 89).

Freeman, Richard and Jo Maybin (2011). "Documents, practices and policy". In: *Evidence & Policy* 7.2. Publisher: Policy Press Section: Evidence & Policy, pp. 155–170. DOI: 10.1332/174426411X579207 (cit. on p. 88).

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences* 120.30. Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.2305016120 (cit. on p. 98).

Gilardi, Fabrizio, Charles R. Shipan, and Bruno Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". In: *American Journal of Political Science* 65.1, pp. 21–35. DOI: 10.1111/ajps.12521 (cit. on p. 88).

Hagendorff, Thilo (2020). "The Ethics of AI Ethics: An Evaluation of Guidelines". In: *Minds and Machines*. DOI: 10.1007/s11023-020-09517-8 (cit. on pp. 87, 89, 90, 94, 114).

Hammitt, James K. et al. (2005). "Precautionary Regulation in Europe and the United States: A Quantitative Comparison". In: *Risk Analysis* 25.5, pp. 1215–1228. DOI: 10.1111/j.1539-6924.2005.00662.x (cit. on p. 93).

Heckelman, Jac C. and Bonnie Wilson (2016). "Interest Groups, Democracy, and Policy Volatility". In: *Contemporary Economic Policy* 34.2, pp. 223–233. DOI: 10.1111/coep.12118 (cit. on p. 90).

Hillman, Amy J., Gerald D. Keim, and Douglas Schuler (2004). "Corporate Political Activity: A Review and Research Agenda". In: *Journal of Management* 30.6. Publisher: SAGE Publications Inc, pp. 837–857. DOI: 10.1016/j.jm.2004.06.003 (cit. on p. 90).

Hine, Emmie and Luciano Floridi (2022). *Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies*. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4006332 (cit. on p. 93).

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9. ZSCC: 0000007, pp. 389–399. DOI: 10.1038/s42256-019-0088-2 (cit. on pp. 89, 94, 102, 114).

John, Peter (2021). "7. Interest Groups, Advocacy, and Policy-Making". In: *British Politics*. Oxford University Press, pp. 209–224. DOI: 10.1093/hepl/9780198840626.003.0007 (cit. on p. 90).

Justo-Hanani, Ronit (2022). "The politics of Artificial Intelligence regulation and governance reform in the European Union". In: *Policy Sciences* 55.1, pp. 137–159. DOI: 10.1007/s11077-022-09452-8 (cit. on p. 93).

Kheiri, Kiana and Hamid Karimi (2023). *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*. arXiv:2307.10234 [cs]. DOI: 10.48550/arXiv.2307.10234 (cit. on p. 98).

Liebig, Laura et al. (2022). "Subnational AI policy: shaping AI in a multi-level governance system". In: *AI & SOCIETY*. DOI: 10.1007/s00146-022-01561-5 (cit. on p. 89).

Macher, Jeffrey T. and John W. Mayo (2015). "Influencing public policymaking: Firm-, industry-, and country-level determinants: Influencing Public Policymaking". In: *Strategic Management Journal* 36.13, pp. 2021–2038. DOI: 10.1002/smj.2326 (cit. on p. 92).

Muddiman, Ashley, Shannon C. McGregor, and Natalie Jomini Stroud (2019). "(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries". In: *Political Communication* 36.2, pp. 214–226. DOI: 10.1080/10584609.2018.1517843 (cit. on p. 98).

Radu, Roxana (2021). "Steering the governance of artificial intelligence: national strategies in perspective". In: *Policy and Society* 40.2, pp. 178–193. DOI: 10.1080/14494035.2021.1929728 (cit. on pp. 89, 90).

Roberts, Huw, Josh Cowls, Emmie Hine, et al. (2021). "Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US". In: *Science and Engineering Ethics* 27.6. DOI: 10.1007/s11948-021-00340-7 (cit. on pp. 87, 91).

Roberts, Huw, Josh Cowls, Jessica Morley, et al. (2021). "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation". In: *AI & SOCIETY* 36.1, pp. 59–77. DOI: 10.1007/s00146-020-00992-2 (cit. on pp. 87, 91).

Roche, Cathy, P. J. Wall, and Dave Lewis (2023). "Ethics and diversity in artificial intelligence policies, strategies and initiatives". In: *AI and Ethics* 3.4, pp. 1095–1115. DOI: 10.1007/s43681-022-00218-9 (cit. on p. 95).

Rodriguez, Juan Cruz (2023). *Interface to 'ChatGPT' from R* (cit. on p. 118).

Schiff, Daniel (2022). "Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy". PhD thesis. DOI: 10.17605/OSF.IO/KW8XD (cit. on p. 90).

Schiff, Daniel, Justin Biddle, et al. (2020). "What's Next for AI Ethics, Policy, and Governance? A Global Overview". In: *Proceedings of the AAAI/ACM Conference on*

AI, Ethics, and Society. New York NY USA: ACM, pp. 153–158. DOI: 10.1145/3375627.3375804 (cit. on p. 89).

Schiff, Daniel, Jason Borenstein, et al. (2021). "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection". In: *IEEE Transactions on Technology and Society* 2.1, pp. 31–42. DOI: 10.1109/TTS.2021.3052127 (cit. on pp. 89, 94, 114).

Schiff, Daniel, Kelly Laas, et al. (2022). "Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies". In: *Codes of Ethics and Ethical Guidelines*. Ed. by Kelly Laas, Michael Davis, and Elisabeth Hildt. Vol. 23. Cham: Springer International Publishing, pp. 121–143 (cit. on p. 88).

Shaffer, Brian (1995). "Firm-level Responses to Government Regulation: Theoretical and Research Approaches". In: *Journal of Management* 21.3, pp. 495–514. DOI: 10.1177/014920639502100305 (cit. on p. 90).

Smuha, Nathalie A. (2019). "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence". In: *Computer Law Review International* 20.4. Publisher: Verlag Dr. Otto Schmidt, pp. 97–106. DOI: 10.9785/cri-2019-200402 (cit. on p. 87).

Tallberg, Jonas, Magnus Lundgren, and Johannes Geith (2023). *AI Regulation in the European Union: Examining Non-State Actor Preferences*. arXiv:2305.11523 [econ, q-fin]. DOI: 10.48550/arXiv.2305.11523 (cit. on pp. 87–89, 92, 94, 109).

Ulnicane, Inga (2022). "Emerging technology for economic competitiveness or societal challenges? Framing purpose in Artificial Intelligence policy". In: *Global Public Policy and Governance* 2.3, pp. 326–345. DOI: 10.1007/s43508-022-00049-8 (cit. on pp. 87, 88).

Ulnicane, Inga et al. (2021). "Framing governance for a contested emerging technology:insights from AI policy". In: *Policy and Society* 40.2, pp. 158–177. DOI: 10.1080/14494035.2020.1855800 (cit. on pp. 87, 88).

Wickham, Hadley (2022). *stringr: Simple, Consistent Wrappers for Common String Operations* (cit. on p. 116).

Zeng, Yi, Enmeng Lu, and Cunqing Huangfu (2018). *Linking Artificial Intelligence Principles* (cit. on p. 89).

3.7 Appendix

3.7.1 Document collection

To collect the texts for this analysis, I initially listed all documents analysed by Fjeld et al. (2020) (n=36), Jobin, Ienca, and Vayena (2019) (n=84), and Schiff, Borenstein, et al. (2021) (n=112). There was some overlap across these collections, resulting in 157 unique documents. By matching these entries with other compilations, notably AlgorithmWatch's "AI Ethics Guidelines Global Inventory" (https://inventory.algorithmwatch.org/) and the OECD AI Policy Observatory (https://oecd.ai), I identified another 43 documents. Since the research papers ended their data collection efforts in 2019 and AlgorithmWatch's inventory has not been updated after April 2020, I conducted additional desk research to identify more recent AI policy documents. This consisted of an extensive online search for relevant keywords, consistent monitoring of relevant news sources and careful study of academic contributions who might refer to relevant documents. In total, this process resulted in the 317 documents taken for this analysis. Appendix 3.7.7 lists these documents together with relevant metadata variables.

To be included in the final collection, a document had to meet a number of criteria, broadly in line with previous collection efforts (e.g., Jobin, Ienca, and Vayena 2019 and Hagendorff 2020): it had to (1) be publicly available in English (for a handful of selected cases, authoritative English translations were used); (2) be published between 2016 and 2022; (3) be issued by an institutional entity from one of the three actor types considered for this study (academic publications or single-authored opinion articles were not included); (4) explicitly mention AI in the title or subtitle ("artificial intelligence" as well as the closely related/synonymous concepts "automated or intelligent systems", "augmented intelligence", "machine learning", but not "algorithmic" or "robotics"); (5) speak to the global level or have significant international influence, either due to the publisher's nature or the document's reception. If newer versions/updates of a document existed, those were listed as separate documents only if there were substantial changes between the versions. Otherwise, only the most recent document has been kept.

When it comes to coding the sector to which a document's publishing author pertained (public sector, business sector, or civil society), the coding process was generally very straightforward and unambiguous. Importantly, these categories are operationalised as mutually exclusive, meaning that an actor can only pertain to one of them. Only for a handful of border cases, the categorisation was more challenging. For instance, some business associations portray themselves as broad coalitions and could arguably be considered to form part of civil society. In those cases, I looked at funding and decision-making structures to determine whether an institution could be considered sufficiently independent from business interests to classify as "civil society". Similarly, there were cases in which an independent government-affiliated body acted as a

watchdog promoting civil society interests. If it was sufficiently evident that this body did indeed have full autonomy from the government, it would be classified as "civil society", otherwise as "public sector". Table 3.2 in the main text provides an overview of how many documents per sector were identified.

Table 3.6 reports the average document length by sector and region, showing that businesses tend to publish the most concise texts whereas public sector documents – especially from Europe – are much longer. There is also considerable variation across regions, especially for documents from the business and public sectors. Notably, the actual average length of public sector documents is even larger, given that many of its texts that exceeded 50 pages had to be truncated. Hence, the table merely serves as an indication of the large variation in document length across sectors and regions, which substantiates that the analysis is focused on relative shares rather than absolute counts.

	public sector	business	civil society
Overall	225.63	81.72	131.96
Europe	269.59	61.62	125.19
North America	169.48	83.10	121.50
Other	200.94	101.71	151.68

Table 3.6: *Average length of documents (number of sentences) by sector and region, after truncating.*

3.7.2 Topic dictionaries and detection accuracy

The dictionaries for each topic were inductively built, starting with keywords and labels commonly found in the literature, and identified by human reading of various documents. I then iteratively fine-tuned the keywords, building more complex regular expressions (for instance, to capture variations of a word, such as plural forms or different spellings) and Boolean operators (to capture prominent word combinations, where each word on its own would not be informative) until they satisfyingly classified a small inspection sample. Building the dictionary for "innovation"-related terms was a relatively straight-forward task, as the term "innovation" and its various variant forms already capture a large part of the relevant sentences. I then added some more specific labels such as "startup" or "sandboxes", which are unlikely to be used in any other context.

The protection dimension turned out to be much harder to identify with a dictionary-based approach. Therefore, I resorted to a nested approach, in which I first detected whether sentences spoke to some of the ethical themes identified by Fjeld et al. (2020). From these themes, I used all but two (promotion of human values, professional responsibility), as they were less clearly tied to the notion of protection that this analysis

is interested in. Then, if a sentence was found to address one of these ethical themes or sub-topics (such as transparency, accountability, or safety), I code it as concerned with "protection". As shown in table 3.8, the accuracy for these classification tasks is very high and on par with human intercoder reliability. Table 3.7 gives a breakdown of the distribution of sub-topics (computed as the average share of sentences per document). "Human control of technology" stands out as the topic with the least mentions (less than 1%).

	Mean share
Privacy	0.06
Accountability	0.08
Safety & security	0.08
Transparency & explainability	0.05
Fairness and non-discrimination	0.09
Human control of technology	0.01
Protection (overall)	0.06

Table 3.7: *Share of sentences (in %) speaking to the various sub-topics of protection.*

The following list includes the dictionaries for innovation as well as the different sub-topics borrowed from Fjeld et al. (2020) that together formed the protection topic. Notably, text was pre-processed (including conversion to lowercase and removal of special characters) before I checked for the presence of the following strings using the *stringr* package (Wickham 2022).

- **Innovation**: *"innovat*; invest*; intellectual property; productivity; startup; incubator; infrastructure; sandboxes; test bed"*
- **Privacy**: *"privacy; personal information; private information; personal data; private data; gdpr; data ownership; data agency; consent; data protection; rectification; erasure; surveill*"*
- **Accountability**: *"(responsible/accountable) * (design/use/development/algorithm/ai); accountab*; liability; (act*) with integrity; impact assessment; evaluation; audit; verify; verifiab*; replicab*; appeal; monitoring; remedy; redress"*
- **Safety and security**: *"nonmaleficence; security; secure; safety; _harm_; harmful; harming; harmed; protection (from/of); attack; function* robust; precaution; prevention; mental integrity; bodily integrity; nonsubversion; reliability; predictability"*
- **Transparency and explainability**: *"(be/are) open about; transparen*; explainab*; opaque; explicab*; traceab*; understandab*; intepretab*; open data; open source data; open algorithm; open government procurement; opacity; disclosure; notification; notify; reporting; right to information"*
- **Fairness and non-discrimination**: *"justice; fairness; (be/are/is) fair; unfair; consistency; inclusion; inclusive; equality; equalit*; equal right; bias; discriminat*; diversity; diverse; plurality; accessibility; benefit * empower; reversibility; remedy; redress; access and distribution; (people/peoples/person/persons) of (color/colour); _minorities_; religious belie*"*
- **Human control**: *"(human/people*/societal) (control/review/oversight/supervision/autonomy); reviewed by human; human (in/on) the loop; opt out; optout; opting out; selfdetermination; transfer (*)*control"*

To measure the accuracy of topic detection via this approach, a sample of 1,000 sentences had been labelled by two human coders independently, and then compared to the computationally assigned labels. One coder was a topic expert, whereas the other was a research assistant without specific knowledge of the topic. To facilitate the human coding process, the online software "Textada" has been used (`https://www.textada.com`). As shown in table 3.8, the dictionary approach performs as well as the human coders. The performance of the computer-based approach yields accuracies between 82% for *innovation* and 98% for *privacy* and *human control of technology*. These were also the tasks where the human coders had the lowest and extremely high agreement, indicating that the scope of *innovation* may simply be harder to define. Overall, the performance of the computer-based approach is very high and in line with the human-to-human intercoder reliability scores. Therefore, I am confident that the approach can be scaled up to the remaining documents and give us reliable information on the topics that each sentence does or does not address.[16]

Theme	Coder 1-coder 2	Machine-coder 1	Machine-coder 2
Privacy	95.9	98.1	95.4
Accountability	94.5	92.8	94.1
Safety & security	92.7	91.1	90.8
Transparency & explainability	95.5	96.1	95.0
Fairness & non-discrimination	94.8	96.6	93.3
Human control of technology	97.2	98.2	97.5
Innovation	86.9.4	86.2	82.4

Table 3.8: *Agreement rate (in per cent) per theme.*

[16] Note that I also experimented with more sophisticated NLP techniques such as pre-trained BERT models to classify sentences into topics. However, even when fine-tuning the classifier on the human-labelled examples, it did not exceed the performance of the dictionary-based approach, so I opted for the more explainable and transparent technique.

3.7.3 Using GPT3.5 to score sentences' stance

Here I report the steps taken to get stance predictions from OpenAI's large language model (LLM) gpt-3.5-turbo. In line with the topic detection approach, text was analysed at the sentence level and only afterwards was the information aggregated to the document level. Notably, I fed the model the sentences without additional pre-processing. This means that casing, stopwords, numbers and punctuation have been kept from the original source text. I used the *chatgpt* package (Rodriguez 2023) to access OpenAI's API via RStudio and prompt it as explained below. Then, I parsed the model responses into a data frame for subsequent analysis.

After experimenting with different settings and model configurations, I settled on a pipeline using the gpt-3.5-turbo-0613 model at a reduced temperature of 0.3. Such a lower temperature forces the model to be less creative in its output, resulting in more stable and predictable responses. I also increased the maximum level of tokens to 1, 024 so that I could feed it longer input.

The prompt I used was the following: *"Does each of the following statements portray a negative or positive view of artificial intelligence (AI)? Negative views of AI may be related to, e.g., technical and ethical risks, how the technology may be misused or harmful, or any other disadvantages of AI. Statements stressing the need for prevention and protection, or improvements in terms of AI quality and trustworthiness, may also indicate negative views of AI. On the other hand, positive views of AI may be related to the opportunities and benefits associated with AI. Based on this information, use a 5-point Likert scale to assess the following statements' views of AI, with 1 being extremely negative, 3 being neutral, and 5 being extremely positive. Explain the reasoning for the score in one sentence. Answer in form of a table that looks like this: —ID—Statement—Score (1-5)—Reasoning—"*

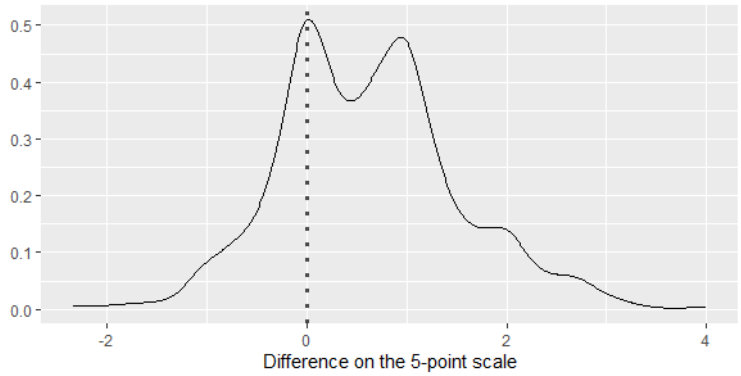
Moreover, I fed batches of input simultaneously, to optimise the speed and resource usage of the model. Generally, these batches would consist of 20 sentences at once, with the model responding in a table format providing scores for the 20 sentences. In cases where the original sentences were extremely long, I decreased the batch size to 10, so as not to exceed the model's maximum token length. To avoid that previous answers might interfere with subsequent predictions, I reset the chat history before each instance.

Accounting for the probabilistic nature of LLMs, I had each sentence analysed independently three times, randomly varying their order to mitigate any concerns about output path dependency. I then took the average score of the three predictions, and also computed the standard deviation to see how reliable the model output was over iterations. In 87% of cases, the three models did not deviate by more than 0.5 points on the 5-point scale. In only 0.7% was the deviation larger than 1 point, indicating a very high level of coherence across iterations.

To gauge the validity and accuracy of the model output, I had a random sample of 500 sentences coded manually by a research assistant, following the same instructions as

were given to the LLM. To assess the degree of agreement between LLM output and human coding, I calculated Cohen's Kappa with quadratic weights (so that more extreme divergences get penalised stronger). The score is 0.503, which is commonly interpreted as moderate agreement. In around 80% of cases was the disagreement smaller than 1 point on the 5-point Likert scale. And encouragingly, grave errors (where the disagreement exceeded 2.5 points, or half the scale), were found in only 1.2% of cases.[17] While not perfect, the validation indicates satisfying levels of reliability to continue with the analysis. But looking at the distribution of the error term also reveals a systematic error bias, with the model leaning towards excessively positive assessments.

Figure 3.4: *Comparing human scores with GPT-3.5-turbo output shows the model's tendency to slightly overestimate sentences' stance.*



Note: The difference is calculated as model_score − human_score.

Figure 3.4 plots the distribution of distance between human and LLM scoring. It shows that the model suggestions are overly positive, with a substantial share of the sentences being rated 1 point higher than the human coder did. To better understand what might be driving this behaviour, I looked at sentences where the mismatch between human- and AI-assigned scores was largest. For instance, the sentence "human beings should remain free to make life decisions for themselves", found in the European Commission's "Ethics Guidelines for Trustworthy AI", has been coded as highly positive (5/5), despite the sentence not allowing for a clear assignment of a stance. Indeed, one could argue that such calls are a reflection of grave concerns over AI's potential threats to human autonomy, and thus the sentence should be scored more negatively. Now,

[17] There was just one incidence in the validation set where the output was on the extremely wrong end of the scale. The model scored the sentence "Viewed through this human rights lens, the co-opted use of an automated system by a bad faith actor creates a human rights liability that demands redress." as 5/5 (very positive), arguing that "This statement portrays a positive view of AI by highlighting the importance of human rights and the need for accountability and redress in AI development." This is clearly not the desired output, but luckily such grave mistakes seem to occur exceedingly rarely.

one of the strengths of using LLMs for coding is that we can also query the model's reasoning for why it assigned a certain score to a given sentence. In this case, it explained the score as follows: "The statement emphasizes the importance of individual freedom, indicating a positive view of AI." A similar misjudgement applies to the sentence "These guidelines are intended to foster responsible and sustainable AI innovation in Europe," again by the European Commission. Here, the model assigned again a score of 5, which the sentence does not actually justify. While fostering AI innovation can generally be seen as reflecting a positive stance towards the technology, the call for "responsible and sustainable" development implies that not all AI is necessarily welcomed. Such misled reasoning is present in many of the excessively positive views and probably explains the overall bias towards higher scores. More episodic evidence of this can also be found in appendix 3.7.4.

Table 3.9 shows the results of the stance prediction, by sector and region. It shows that estimates are remarkably stable across regions. Differences are more obvious when comparing different sectors, with businesses being the most positive and civil society the most negative actors. In the main text, I present changes to stance over time, where some noteworthy variation occurs.

	public sector	business	civil society
Overall	3.81	4.06	3.50
Europe	3.81	4.08	3.44
North America	3.74	4.03	3.45
Other	3.84	4.05	3.64

Table 3.9: *Mean stance of documents. By sector and region.*

3.7.4 Example output of the computational text analysis

sentence_id	doc_id	sentence	prot.	innov.	stance
6750	1042	This technology is already being used for discriminatory and unethical purposes, often without people's knowledge.	1	0	1.00
769	1000	Any form of citizen scoring can lead to the loss of this autonomy and endanger the principle of non-discrimination.	1	0	1.33
6397	1040	This is why understanding AI systems' potential impact on labor is an important aspect of understanding their impact on economic equality, and preparing accordingly.	1	0	1.33
7739	1057	We recognize that AI technologies can help promote inclusive economic growth, bring great benefits to society, and empower individuals.	1	0	5.00
22673	1157	Harnessed appropriately, we believe AI can deliver great benefits for economies and society, and support decision-making which is fairer, safer and more inclusive and informed.	1	0	5.00
198	1000	Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process.	1	0	4.67
6822	1043	Across Asia, Africa, and Latin America, multiple governments continue to roll out biometric ID projects that create the infrastructure for both state and commercial surveillance.	1	1	1.33
7818	1057	In order to foster an open, transparent and conducive global policy environment for investment, we recognize the value of improving open, non-discriminatory, transparent and predictable conditions for investment.	1	1	3.33
51136	1326	Incorporating AI as a factor to improve the productivity of Spanish business, to increase the efficiency of public administration, and to drive sustainable and inclusive economic growth.	1	1	5.00
48956	1303	Establishing clear, innovation-friendly and flexible approaches to regulating AI will be core to achieving our ambition to unleash growth and innovation while safeguarding our fundamental values and keeping people safe and secure.	1	1	4.67
445	1000	Moreover, sometimes small changes in data values might result in dramatic changes in interpretation, leading the system to e.g. confuse a school bus with an ostrich.	0	0	1.33
52676	1339	The sequence of instructions that a computer uses to predict a person's age (output) from their picture (input) is also defined by an algorithm.	0	0	3.00
33334	1211	Developing AI-based solutions will also provide an opportunity to demonstrate the application and potential of AI technologies, encouraging greater adoption across the economy.	0	0	5.00
38673	1232	Singaporeans will enjoy greater safety and security through our strengthened border security.	1	0	4.67

Table 3.10: *Example of sentences and their assigned labels and scores. Faulty assignments are highlighted in red.*

Table 3.10 presents a purposefully drawn sample of sentences across various documents, with the goal of illustrating instances in which the automated text analysis approach returns incorrect assignments. In red, there are two instances in which the stance scores produced by the LLM diverged from the human reviewer. First, for sentence 6397, the model gave an excessively negative score to the sentence. While it does mention the "potential impact on labor" and, by extension, on economic quality, it is not specified whether this impact is necessarily negative. Hence, the very low score of

1.33 seems a bit extreme. In turn, sentence 198 has received a very high score of 4.67, despite it conveying a cautionary message. By expressing the need for humans to maintain self-determination when interacting with AI systems, the sentence suggests a certain degree of concern, which is not reflected accurately in the positive score. As mentioned in appendix 3.7.3, the model in general is slightly biased towards such excessively positive scores, as it often mistakenly interprets intentions as descriptions or factual statements, whereas in reality they often reflect an underlying normative concern of what ought to be.

In blue are marked instances in which the topic detection approach has picked up wrong information. Sentence 51136 is a false positive for the protection topic, triggered by the word "inclusive." While the statement's concern with inclusive growth arguably reflects a certain concern with the protection of individual and collective rights, this aspect is not directly linked to the actual AI system. Sentence 33334, at last, is an example of a false negative. It is clearly indicative of an innovation-friendly view, but has not been picked up by the innovation dictionary.

3.7.5 Additional tables and figures

Robustness check: topic attention and document length

Since the distribution of sentences per topic – reported in 3.4 – may be strongly affected by document length, I offer an alternative computation as a robustness check. 3.11 reports regions' and sectors' mean attention, with numbers first aggregated at the document-level to account for differences in document length. The overall pattern (i.e., the relative position of regions and actors) remains the same.

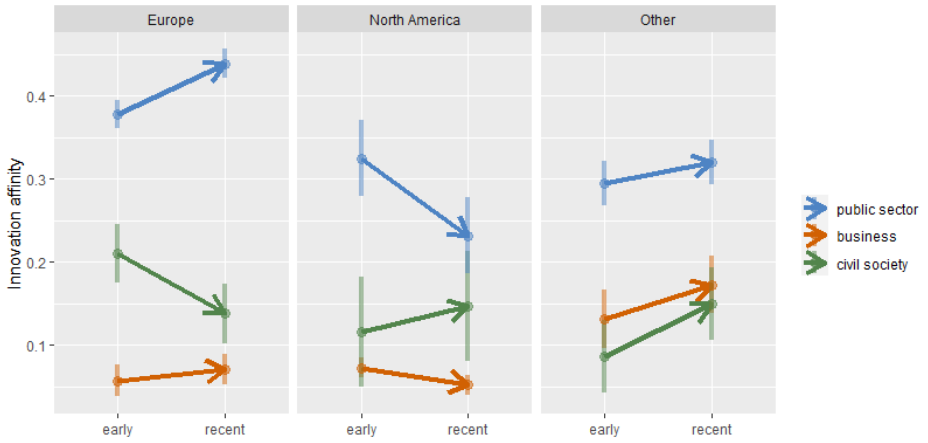
	(1) Innovation	(2) Protection	(5) Innovation affinity
Overall	7.3	28.4	20.5
public sector	10.5	22.9	31.5
Business	3.3	34.8	8.6
Civil society	4.7	33.4	12.3
Europe	8.2	24.1	25.4
North America	5.8	36.0	13.8
Other	7.5	27.4	21.4

Table 3.11: *Share of sentences (in %) speaking to innovation, protection, both, or none, as well as innovation affinity. Average scores by sector and region, first aggregated at document level.*

Robustness check: including the "Other" region

Running the analysis for the full sample (with the region of remaining – i.e., neither American nor European – actors coded as "Other"), the public sector remains on the top regarding the attention to innovation (figure 3.5). Unlike for the transatlantic actors, businesses are taking over civil society. As expected, there is a less clear picture when it comes to the public sector's shift in preferences, as different dynamics in different sub-regions probably cancel each other out to a considerable extent. While the direction for civil society seems to point clearly towards more innovation relative to protection, this difference is also not statistically significant.

Figure 3.5: *Mean attention to innovation versus protection, by sector, region, and over time.*



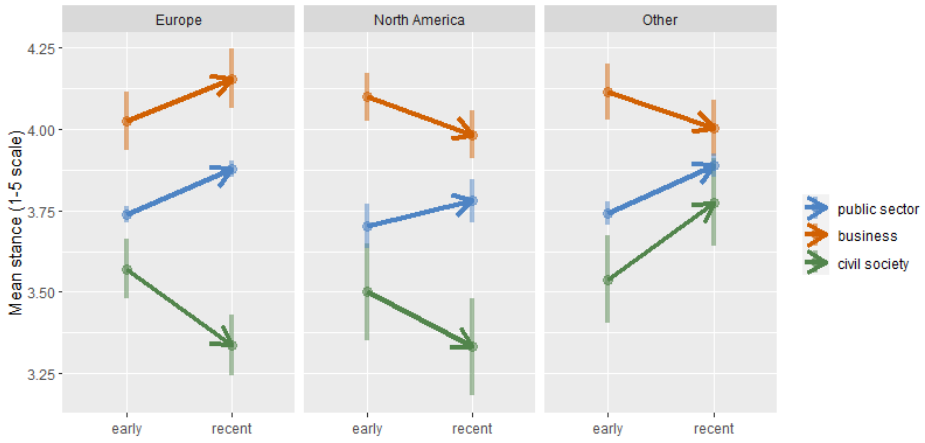
Note: Innovation affinity reports the share of sentences on innovation versus sentences on innovation and/or protection. A value of 0 implies that relevant sentences in a document have exclusively covered protection, whereas a value of 1 implies that they have exclusively covered innovation. Error bars indicate the 95% confidence intervals for differences between two time periods.

3.7.6 Limitations

This paper puts forward a novel technique for the analysis of AI policy documents. At the aggregate level, the findings of this sentence-based analysis are largely in line with previous research, which gives confidence that the dictionary-based topic detection approach is robust. Nevertheless, there are important limitations and caveats to consider.

First of all, dictionary-based approaches suffer from several shortcomings. They are not able to interpret context and may struggle with words that have competing, context-specific meanings. They may also fail to pick up on some rarely used terms or analogies, which the researcher did not include in the dictionary. Still, for this use case it provided a sufficiently reliable measure which performs equally accurate to human coders and more advanced machine-learning techniques. As the comparison to human reviewers has shown, error rates range from between 5 to 20 %, depending on the theme. This indicates that some themes are more easy to detect based on keywords than others. On the upside, it is plausible to assume that the error rates should be more or less evenly distributed across documents and sectors, so that misidentifications even out at the aggregate level. That is also why this approach leads to robust results when aggregated, even when it cannot give reliable information for individual sentences. One concern here may be that some shorter documents are more vulnerable to grave errors due to misclassification. In other words, if a sentence gets misclassified in a document that consists of only ten statements, the effect on aggregate shares will be very strong. For

Figure 3.6: *Mean stance. By sector, region, and over time.*



Note: Stance based on averaged GPT3.5-turbo predictions. Error bars indicate the 95% confidence intervals for differences between two time periods.

longer documents, such a misclassification should have relatively less weight and therefore be much less problematic. To account for this, the shortest documents (with less than 15 sentences) have all been manually reviewed by the author. Where applicable, the human corrections supersede the machine predictions. Encouragingly, across the 15 shortest documents, corrections were necessary in only 18% of cases for stance, and only 8% of all possible labelling decisions. This higher accuracy for shorter documents may to be a function of authors' need to express themselves in clear and unambiguous terms, which greatly improves the functioning of the dictionary lookup. Second of all, while sentence-based theme information adds a great amount of nuance to the discussion, it is still not a perfect representation of authors' priorities. Even if an author dedicates only a handful of sentences to a specific theme, these sentences may be so strongly written that they express a much higher prioritisation than another theme where authors waver and take up several paragraphs without saying anything meaningful. Such differences in the quality, intent, or direction of sentences cannot be measured with this paper's method, which only captures raw frequencies of attention.

The usage of an LLM's output as a measurement for sentences' stance also presents some challenges that should be acknowledged. For one, the underlying LLM belongs to a private company (OpenAI), so possible future replications of this work will always be contingent on that company's continued provision of the model and necessary API services. Moreover, as the company has decided to not release an open-source version (at least for now), the research community has limited means to understand and investigate the model's behaviour and explain the output. This makes it harder to detect potential biases or other flaws that may put in question the model's output.

A last aspect concerns the document collection process, which may have missed some important documents while unduly including some others. Language is one barrier, the author's geographic and epistemic constraints are another. While great efforts have been taken to make sure that the collection is as comprehensive and diverse as possible, it cannot claim to be complete. And in turn, while specific criteria have been developed and applied to ensure that only relevant documents were included, these are nevertheless reliant on subjective interpretations by the researcher (e.g., when does a company blog post on AI policy become a sufficiently meaningful publication to be included in the analysis?).

3.7.7 List of all documents

Table 3.12 reports the 317 documents that were included in the analysis, as well as relevant metadata variables.

Document	Year	Region	Sector
The Greens (Green Working Group Robots): Position on Robotics and Artificial Intelligence	2016	Europe	civil society
United Kingdom: House of Commons, Science and Technology Committee: Robotics and artificial intelligence	2016	Europe	public sector
Accenture Labs: Universal Principles of Data Ethics	2016	Global	business
Partnership On AI (Apple, Amazon, Google, MS, etc): Tenets	2016	North America	business
Fairness, Accountability, and Transparency in Machine Learning (FATML): Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	2016	North America	civil society
AI Now Institute: The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	2016	North America	civil society
Executive Office of the President; National Science and Technology Council; Committee on Technology: Preparing for the future of Artificial Intelligence	2016	North America	public sector
National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee: The National Artificial Intelligence Research and Development Strategic Plan	2016	North America	public sector
United States: Executive Office of the President: Artificial Intelligence, Automation, and the Economy	2016	North America	public sector
Government of UAE: National Strategy for Artificial Intelligence	2017	Africa and Middle East	public sector
Japanese Society for Artificial Intelligence: The Japanese Society for Artificial Intelligence Ethical Guidelines	2017	Asia-Pacific	civil society
Chinese government: New Generation Artificial Intelligence Development Plan	2017	Asia-Pacific	public sector
Advisory Board on Artificial Intelligence and Human Society (initiative of the Minister of State for Science and Technology Policy): Report on Artificial Intelligence and Human Society (Unofficial translation)	2017	Asia-Pacific	public sector
Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society: Draft AI R&D Guidelines for International Discussions	2017	Asia-Pacific	public sector
National Information Society Agency: Ethical Issues of Artificial Intelligence using Future Signal Detection Methods	2017	Asia-Pacific	public sector
National Information Society Agency: Artificial Intelligence and the Future of Jobs	2017	Asia-Pacific	public sector
Sage: The Ethics of Code: Developing AI for Business with Five Core Principles	2017	Europe	business
DeepMind: DeepMind Ethics & Society Principles	2017	Europe	business
Bitkom: Artificial Intelligence	2017	Europe	business
The Royal Society: Machine learning: the power and promise of computers that learn by example	2017	Europe	civil society
UNI Global Union: 10 Principles for Ethical AI	2017	Europe	civil society
French Data Protection Authority (CNIL): How can humans keep the upper hand?	2017	Europe	public sector
Information Commissioner's Office: Big data, artificial intelligence, machine learning and data protection	2017	Europe	public sector

Document	Year	Region	Sector
Finnish Ministry of Economic Affairs and Employment: Finland's Age of Artificial Intelligence: Turning Finland into a leading country in the application of artificial intelligence, Objective and recommendations for measures	2017	Europe	public sector
European Economic and Social Committee (EESC): Artificial Intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment, and society	2017	Europe	public sector
France: Parliamentary Office for the Evaluation of Scientific and Technological Choices (OPECST): Toward a Controlled, Useful and Demystified Artificial Intelligence	2017	Europe	public sector
United Kingdom: All Party Parliamentary Group on AI (APPG AI) Secretariat: APPG AI Findings 2017	2017	Europe	public sector
Information Technology Industry Council (ITI): AI Policy Principles	2017	Global	business
Software & Information Industry Association (SIIA), Public Policy Division: Ethical Principles for Artificial Intelligence and Data Analytics	2017	Global	business
UNI Global Union: Top 10 Principles for Ethical AI	2017	Global	civil society
Internet Society: Artificial Intelligence and Machine Learning: Policy Paper	2017	Global	civil society
COMEST/UNESCO: Report of COMEST on Robotics Ethics	2017	Global	public sector
Microsoft: Microsoft AI Principles	2017	North America	business
Intel: Artificial Intelligence: The Public Policy Opportunity	2017	North America	business
Future of Life Institute: Asilomar AI Principles	2017	North America	civil society
Association for Computing Machinery (ACM): Statement on Algorithmic Transparency and Accountability	2017	North America	civil society
AI Now Institute: AI Now 2017 Report	2017	North America	civil society
The Future Society: Principles for the Governance of AI	2017	North America	civil society
The Information Accountability Foundation: Artificial Intelligence, Ethics and Enhanced Data Stewardship	2017	North America	civil society
United States: Department of Homeland Security, Office of Cyber and Infrastructure Analysis: Artificial Intelligence Risk to Critical Infrastructure	2017	North America	public sector
Smart Dubai: AI Ethics Principles & Guidelines	2018	Africa and Middle East	public sector
Kakao Corp: AI Ethics	2018	Asia-Pacific	business
Sony: AI Engagement within Sony Group	2018	Asia-Pacific	business
Baidu: Four Principles of AI Ethics	2018	Asia-Pacific	business
Tencent: "ARCC": An Ethical Framework for Artificial Intelligence	2018	Asia-Pacific	business
The Centre for Internet & Society: Artificial Intelligence in the Governance Sector in India	2018	Asia-Pacific	civil society
Standards Administration of China: White Paper on AI Standardization	2018	Asia-Pacific	public sector
Monetary Authority of Singapore: Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector	2018	Asia-Pacific	public sector
National Institution for Transforming India (Niti Aayog): Discussion Paper: National Strategy for Artificial Intelligence	2018	Asia-Pacific	public sector
Chinese Academy of Sciences, Institute of Automation, Research Center for Brain-inspired Intelligence: Harmonious Artificial Intelligence Principles	2018	Asia-Pacific	public sector
India: NITI Aayog: National Strategy for Artificial Intelligence: #AIForAll	2018	Asia-Pacific	public sector
Government of Korea: Ethics Guideline for the Intelligent Information Society	2018	Asia-Pacific	public sector
Personal Data Protection Commission Singapore: Discussion Paper on Artificial Intelligence (AI) and Personal Data -Fostering Responsible Development and Adoption of AI	2018	Asia-Pacific	public sector
Government of Korea: Ethics Charter for the Intelligent Information Society	2018	Asia-Pacific	public sector
SAP: SAP's guiding principles for artificial intelligence	2018	Europe	business
Deutsche Telekom: Guidelines for Artificial Intelligence	2018	Europe	business

Document	Year	Region	Sector
Institute of Business Ethics: Business Ethics and Artificial Intelligence	2018	Europe	business
OP Group: OP Financial Group's ethical guidelines for artificial intelligence	2018	Europe	business
Digital Catapult: Machine Intelligence Garage Ethics Committee: Machine Intelligence Garage Ethics Framework	2018	Europe	business
Futurice: The Futurice Principles for Ethical AI	2018	Europe	business
Phrasee: Phrasee's AI Ethics Policy	2018	Europe	business
Telefónica: Telefónica´s Approach to the Responsible Use of AI	2018	Europe	business
Tieto: Tieto's AI ethics guidelines	2018	Europe	business
Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	2018	Europe	civil society
Biocat and l'Obra Social la Caixa: The Barcelona declaration for the proper development and usage of artificial intelligence in Europe	2018	Europe	civil society
European Commission, The European Group on Ethics in Science and New Technologies (EGE): Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	2018	Europe	civil society
AI4People: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	2018	Europe	civil society
IPN Special Interest Group on Artificial Intelligence: Dutch Artificial Intelligence Manifesto	2018	Europe	civil society
The Institute for Ethical AI & Machine Learning: The Responsible Machine Learning Principles: A practical framework to develop AI responsibly	2018	Europe	civil society
Nesta: 10 principles for public sector use of algorithmic decision making	2018	Europe	civil society
Key recommendations: Artificial Intelligence (AI) in Health	2018	Europe	civil society
AI for Humanity, Government of France: For a Meaningful Artificial Intelligence. Towards a French and European strategy	2018	Europe	public sector
UK House of Lords, Select Committee on Artificial Intelligence: AI in the UK	2018	Europe	public sector
European Commission: Artificial Intelligence for Europe	2018	Europe	public sector
Nordic Council of Ministers, Government of Sweden: AI in the Nordic-Baltic region	2018	Europe	public sector
Council of Europe: European Commission for the efficiency of Justice (CEPEJ): European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment	2018	Europe	public sector
German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs: National Strategy for Artificial Intelligence: AI Made in Germany	2018	Europe	public sector
Austria: Austrian Council on Robotics and Artificial Intelligence, Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology: Shaping the Future of Austria with Robotics and Artificial Intelligence	2018	Europe	public sector
Government of Austria: Artificial Intelligence Mission Austria 2030	2018	Europe	public sector
European Commission, EU Member States: Declaration of Cooperation on AI	2018	Europe	public sector
Ministry of Economic Affairs and Employment: Work in the age of artificial intelligence. Four perspectives on the economy, employment, skills and ethics (only section "Good application of artificial intelligence technology and ethics")	2018	Europe	public sector
France: CERNA: The ethics of research in machine learning	2018	Europe	public sector
Italy: Task Force on Artificial Intelligence of the Agency for Digital Italy (AGID) and Department of Public Administration: Artificial Intelligence at the service of citizens	2018	Europe	public sector
United Kingdom: All Party Parliamentary Group on AI (APPG AI) Secretariat: APPG AI Findings 2018	2018	Europe	public sector
Vinnova: Artificial intelligence in Swedish business and society	2018	Europe	public sector

Document	Year	Region	Sector
The Norwegian Data Protection Authority: Artificial intelligence and privacy	2018	Europe	public sector
World Government Summit: Global Governance of AI Roundtable	2018	Global	civil society
Access Now: Human rights in the Age of AI	2018	Global	civil society
The Public Voice: Universal Guidelines for Artificial Intelligence	2018	Global	civil society
W20: Artificial Intelligence: open questions about gender inclusion	2018	Global	civil society
Privacy International & Article 19: Privacy and Freedom of Expression In the Age of Artificial Intelligence	2018	Global	civil society
WEF, Global Future Council on Human Rights 2016-2018: White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	2018	Global	civil society
International Conference of Data Protection and Privacy Commissioners (ICDPPC): Declaration on ethics and data protection in Artificial Intelligence	2018	Global	public sector
Leaders of the G7: Charlevoix Common Vision for the Future of Artificial Intelligence	2018	Global	public sector
IBM: IBM's Principles for Trust and Transparency	2018	North America	business
IBM: IBM Everyday Ethics for AI	2018	North America	business
Intel Corporation: Intel's AI Privacy Policy White Paper. Protecting individuals' privacy and data in the artificial intelligence world	2018	North America	business
Google: AI at Google: Our Principles	2018	North America	business
Microsoft: The Future Computed – Artificial intelligence and its role in society	2018	North America	business
OpenAI: OpenAI Charter	2018	North America	business
Unity Technologies: Introducing Unity's Guiding Principles for Ethical AI	2018	North America	business
GE Healthcare: AI Principles	2018	North America	business
Microsoft: Responsible bots: 10 guidelines for developers of conversational AI	2018	North America	business
Université de Montréal: Montréal Declaration: Responsible AI	2018	North America	civil society
Data & Society: Governing Artificial Intelligence. Upholding Human Rights & Dignity	2018	North America	civil society
AI Now Institute: AI Now 2018 Report	2018	North America	civil society
Access Now; Amnesty International: The Toronto Declaration: Protecting the right to equality and nondiscrimination in machine learning systems	2018	North America	civil society
United States: Center for Strategic & International Studies (CSIS): A National Machine Intelligence Strategy for the United States	2018	North America	civil society
American Medical Association (AMA): Policy Recommendations on Augmented Intelligence in Health Care H-480.840	2018	North America	civil society
CIGI Gentre for International Governance Innovation: Toward a G20 Framework for Artificial Intelligence in the Workplace	2018	North America	civil society
Electronic Frontier Foundation (EFF): The Cautious Path to Strategic Advantage: How Militaries Should Plan for AI	2018	North America	civil society
United States: Subcommittee on Information Technology, Committe on Oversight and Government Reform: Rise of the Machines: Artificial Intelligence and its Growing Impact on U.S. Policy	2018	North America	public sector
Mexico: British Embassy in Mexico, Oxford Insights, and C Minds: Towards an AI Strategy in Mexico: Harnessing the AI Revolution	2018	North America	public sector
Treasury Board of Canada Secretariat: Responsible use of artificial intelligence	2018	North America	public sector
Government of Mexico: Inteligencia Artificial MX 2018	2018	North America	public sector
Qatar: Qatar Center for Artificial Intelligence (QCAI), Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University: Blueprint: National Artificial Intelligence Strategy for Qatar	2019	Africa and Middle East	public sector
Fujitsu Limited: Fujitsu Group AI Commitment	2019	Asia-Pacific	business
Artificial Intelligence Industry Alliance (AIIA): Joint Pledge on Artificial Intelligence Industry Self-Discipline (Draft for Comment)	2019	Asia-Pacific	business
NEC: NEC Group AI and Human Rights Principles	2019	Asia-Pacific	business
NTT: NTT DATA Group's AI Guidelines	2019	Asia-Pacific	business
Megvii: Core Principles of its AI Practice	2019	Asia-Pacific	business

Document	Year	Region	Sector
Samsung AI Principles	2019	Asia-Pacific	business
Beijng Academy of AI: Beijing AI Principles	2019	Asia-Pacific	civil society
Korea Artificial Intelligence Ethics Association: The AI Ethics Charter	2019	Asia-Pacific	civil society
New Zealand: AI Forum of New Zealand: Artificial Intelligence: Shaping a Future New Zealand: An Analysis of the Potential Impact and Opportunity of Artificial Intelligence on New Zealand's Society and Economy	2019	Asia-Pacific	civil society
Shanghai Declaration of Chinese Young Scientists 2019 Artificial Intelligence Innovation Governance	2019	Asia-Pacific	civil society
Government of Japan; Cabinet Office; Council for Science, Technology and Innovation: Social Principles of Human-Centric AI	2019	Asia-Pacific	public sector
Government of Australia: AI Ethics Principles	2019	Asia-Pacific	public sector
Chinese National Governance Committee for AI: Governance Principles for a New Generation of AI	2019	Asia-Pacific	public sector
Department of Industry Innovation and Science: Artificial Intelligence. Australia's Ethics Framework. A discussion Paper	2019	Asia-Pacific	public sector
Integrated Innovation Strategy Promotion Council (Cabinet Office): AI Strategy 2019	2019	Asia-Pacific	public sector
Government of Russia: National Strategy for the Development of Artificial Intelligence	2019	Asia-Pacific	public sector
Government of Korea: National Strategy for Artificial Intelligence	2019	Asia-Pacific	public sector
Smart Nation and Digital Government Office; National AI Office: National Artificial Intelligence Strategy	2019	Asia-Pacific	public sector
UNESCO, Ministry of Education of the Peoples Republic of China: BEIJING CONSENSUS on artificial intelligence and education	2019	Asia-Pacific	public sector
Mercedes Benz Group: Two Letters and Four Principles: How Mercedes-Benz Uses Artificial Intelligence (AI)	2019	Europe	business
Future Advocacy: Ethical, social, and political challenges of Artificial Intelligence in Health	2019	Europe	business
Telia Company: Guiding Principles on Trusted AI Ethics	2019	Europe	business
PriceWaterhouseCoopers UK: The responsible AI framework	2019	Europe	business
Vodafone: Artificial Intelligence Framework	2019	Europe	business
Philips: Philips AI Principles	2019	Europe	business
Women Leading in AI: 10 Principles of Responsible AI	2019	Europe	civil society
The Public Voice: Declaration: A Moratorium on Facial Recognition Technology for Mass Surveillance	2019	Europe	civil society
The Alan Turing Institute: Understanding artificial intelligence ethics and safety	2019	Europe	civil society
ITechLaw: Responsible AI Policy Framework	2019	Europe	civil society
Bertelsmann Stiftung & iRights.lab:: AlgoRules	2019	Europe	civil society
High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI	2019	Europe	public sector
UK Government: A guide to using Artificial Intelligence in the public sector	2019	Europe	public sector
French Ministry of Defence: ARTIFICIAL INTELLIGENCE IN SUPPORT OF DEFENCE	2019	Europe	public sector
Sweden: Ministry of Enterprise and Innovation: National approach for artificial intelligence	2019	Europe	public sector
Government of Lithuania: LITHUANIAN ARTIFICIAL INTELLIGENCE STRATEGY	2019	Europe	public sector
Malta: Parliametary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: Malta: Towards An AI Strategy	2019	Europe	public sector
Spain: Ministry of Science, Innovation and Universities: Spanish RDI Strategy in Artificial Intelligence	2019	Europe	public sector
Government of Denmark: National Strategy for Artificial Intelligence	2019	Europe	public sector
Conference of the Independent Federal and State Data Protection Supervisory Authorities of Germany: Hambach Declaration on Artificial Intelligence	2019	Europe	public sector

Document	Year	Region	Sector
Lithuania: Ministry of the Economy and Innovation: Lithuanian Artificial Intelligence Strategy: A Vision of the Future	2019	Europe	public sector
Government of Belgium: AI4Belgium Report	2019	Europe	public sector
Government of Czechia: National Artificial Intelligence Strategy of the Czech Republic	2019	Europe	public sector
Government of Luxembourg: Artificial Intelligence: a strategic vision for Luxembourg.	2019	Europe	public sector
Government of Estonia: Report of Estonia's AI Taskforce	2019	Europe	public sector
Government of Portugal: AI PORTUGAL 2030 - NATIONAL STRATEGY FOR AI	2019	Europe	public sector
Government of Finland: Leading the way into the age of artificial intelligence	2019	Europe	public sector
Government of Estonia: Estonia's national artificial intelligence strategy 2019-2021	2019	Europe	public sector
Government of Malta: Malta - The Ultimate AI Launchpad	2019	Europe	public sector
Government of The Netherlands: Strategic Action Plan for Artificial Intelligence	2019	Europe	public sector
Government of Romania: Romania in the era of Artificial Intelligence	2019	Europe	public sector
Catalan regional government: Catalonia's Artificial Intelligence Strategy	2019	Europe	public sector
Swiss Academy of Engineering Sciences: Recommendations for an AI Strategy in Switzerland	2019	Europe	public sector
OECD: OECD Principles on Artificial Intelligence	2019	Europe	public sector
European Commission: The future of work? Work of the future!	2019	Europe	public sector
European Parliament: A comprehensive European industrial policy on artificial intelligence and robotics	2019	Europe	public sector
Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design, First Edition (EAD1e)	2019	Global	business
World Government Summit: AI Ethics: The Next Big Thing in Government	2019	Global	civil society
Conclusion: Ethics of AI in Radiology: European and North American Multisociety Statement	2019	Global	civil society
OECD: Recommendation of the Council on Artificial Intelligence	2019	Global	public sector
G20: G20 AI Principles	2019	Global	public sector
Accenture: Responsible AI: A Framework for Building Trust in Your AI Solutions	2019	North America	business
Google: Perspectives on Issues in AI Governance	2019	North America	business
Workday: Workday's Commitments to Ethical AI	2019	North America	business
Thomson Reuters: Our AI principles	2019	North America	business
Salesforce: Trusted AI Commitment	2019	North America	business
New Work Summit: Seeking Ground Rules for AI	2019	North America	civil society
AI Now Institute: AI Now 2019 Report	2019	North America	civil society
American Association of Artificial Intelligence (AAAI) with Computing Community Consortium (CCC): A 20-Year Community Roadmap for Artificial Intelligence Research in the US	2019	North America	civil society
World Economic Forum: AI Governance: A Holistic Approach to Implement Ethics into AI	2019	North America	civil society
American Association of Artificial Intelligence (AAAI): AAAI Code of Professional Ethics and Conduct	2019	North America	civil society
Treasury Board of Canada Secretariat: Responsible Artificial Intelligence in the Government of Canada (Digital Disruption White Paper Series)	2019	North America	public sector
National Research Council Canada: Advisory Statement on Human Ethics in Artificial Intelligence and Big Data Research (2017)	2019	North America	public sector
US Government: Executive Order 13859	2019	North America	public sector
Canada: Innovation, Science and Economic Development Canada: Declaration of the International Panel on Artificial Intelligence	2019	North America	public sector
United States: Department of Defense: Summary of the 2018 Department Of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity	2019	North America	public sector

Defense Innovation Board: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense	2019	North America	public sector
Office of the Director of National Intelligence: The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines	2019	North America	public sector
National Security Commission on Artificial Intelligence: Interim Report	2019	North America	public sector
IA Latam: Ética IA Latam	2019	South America	civil society
Government of Uruguay: Artificial Intelligence Strategy for the Digital Government	2019	South America	public sector
Government of Argentina: Plan Nacional de Inteligencia Artificial	2019	South America	public sector
Research ICT Africa: An African perspective on gender and artificial intelligence needs African data and research	2020	Africa and Middle East	civil society
Research ICT Africa: The public-private: a key legal nexus for South Africa's AI future	2020	Africa and Middle East	civil society
Government of Israel (PMO Office): THE NATIONAL INITIATIVE FOR SECURED INTELLIGENT SYSTEMS	2020	Africa and Middle East	public sector
Government of Saudi Arabia: National Strategy for Data & AI	2020	Africa and Middle East	public sector
SenseTime: Code of Ethics for AI Sustainable Development	2020	Asia-Pacific	business
AI Forum New Zealand: Trustworthy AI in Aotearoa	2020	Asia-Pacific	civil society
Government of Australia: An AI Action Plan for all Australians	2020	Asia-Pacific	public sector
Personal Data Protection Commission Singapore: MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK	2020	Asia-Pacific	public sector
Government of The Philippines: Artificial Intelligence and Information & Communications Technology	2020	Asia-Pacific	public sector
National Information Society Agency: Policy Tasks for the Stewardship of Trustworthy AI	2020	Asia-Pacific	public sector
National Information Society Agency: Practical Guide for the Stewardship of Trustworthy AI in Public Organizations	2020	Asia-Pacific	public sector
Novartis: Novartis' commitment to the ethical and responsible use of Artificial Intelligence (AI) Systems	2020	Europe	business
Bosch: Bosch code of ethics for AI	2020	Europe	business
AI Now Institute, City of Amsterdam, City of Helsinki, Mozilla Foundation and Nesta: Using procurement instruments to ensure trustworthy AI	2020	Europe	civil society
AI Ethics Impact Group: From Principles to Practice	2020	Europe	civil society
TUM School of Governance: White Paper on AI Ethics and Governance	2020	Europe	civil society
Institute of Informatics and Telecommunications (IIT): Democratising AI - A National Strategy for Greece	2020	Europe	civil society
Rome Call for Ethics: Vatican AI Principles	2020	Europe	civil society
European Commission: On Artificial Intelligence - A European approach to excellence and trust	2020	Europe	public sector
Council of Europe: Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems	2020	Europe	public sector
European Council: Presidency conclusions on the charter of fundamental rights in the context of artificial intelligence and digital change	2020	Europe	public sector
Government of Germany: Artificial Intelligence Strategy of the German Federal Government	2020	Europe	public sector
Government of Germany: Executive Summary of the Final Report of the Study Commission on Artificial Intelligence	2020	Europe	public sector
Government of Serbia: Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025	2020	Europe	public sector
Government of Switzerland: Guidelines on Artificial Intelligence for the Confederation	2020	Europe	public sector
Government of Norway: The National Strategy for Artificial Intelligence	2020	Europe	public sector
Government of Hungary: Hungary's Artificial Intelligence Strategy	2020	Europe	public sector
Government of Bulgaria: CONCEPT FOR THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE IN BULGARIA UNTIL 2030	2020	Europe	public sector

Government of Poland: Policy for the Development of Artificial Intelligence in Poland from 2020	2020	Europe	public sector
Government of Switzerland: Specific Guidelines for the policy area Education, Research and Innovation	2020	Europe	public sector
Government of Spain: National Artificial Intelligence Strategy	2020	Europe	public sector
World Economic Forum: AI Procurement in a Box: Workbook	2020	Global	business
GLOBAL PRIVACY ASSEMBLY: ADOPTED RESOLUTION ON ACCOUNTABILITY IN THE DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE	2020	Global	civil society
NA: Joint Statement from founding members of the Global Partnership on Artificial Intelligence	2020	Global	public sector
AI4D: AI in Africa Policy Project – AI4D	2020	Global	public sector
National Association of Insurance Commissioners (NAIC): National Association of Insurance Commissioners (NAIC) Principles on Artificial Intelligence (AI)	2020	North America	business
Algorithmic Justice League: THE ALGORITHMIC JUSTICE LEAGUE'S 101 OVERVIEW	2020	North America	civil society
Algorithmic Justice League: The Algorithmic Justice League's 101 Overview for equitable and accountable AI	2020	North America	civil society
UK and US governments: Declaration of the United States of America and the United Kingdom of Great Britain and Northern Ireland on Cooperation in Artificial Intelligence Research and Development: A Shared Vision for Driving Technological Breakthroughs in Artificial Intelligence	2020	North America	public sector
US Government: Executive Order 13960	2020	North America	public sector
US Government: American AI Initiative: Year One Annual Report	2020	North America	public sector
US Government: MEMORANDUM - Guidance for Regulation of Artificial Intelligence Applications	2020	North America	public sector
NA: Agenda Nacional Mexicana de Inteligencia Artificial	2020	North America	public sector
U.S. Department of Defense: Ethical Principles for Artificial Intelligence	2020	North America	public sector
Office of the Director of National Intelligence: Principles of Artificial Intelligence Ethics for the Intelligence Community	2020	North America	public sector
Government of Egypt: EGYPT NATIONAL AI STRATEGY	2021	Africa and Middle East	public sector
Naver: AI Ethics Principles	2021	Asia-Pacific	business
AI Alliance Russia: AI Ethics Code	2021	Asia-Pacific	business
SenseTime: AI Ethics for Balanced Development	2021	Asia-Pacific	business
National Governance Committee for the New Generation Artificial Intelligence, Ministry of Science and Technology: Ethical Norms for New Generation Artificial Intelligence	2021	Asia-Pacific	public sector
China Academy of Information and Communications Technology: White Paper on Trustworthy Artificial Intelligence	2021	Asia-Pacific	public sector
India: NITI Aayog: Approach Document for India	2021	Asia-Pacific	public sector
Ministry of Economy, Trade and Industry: Governance Guidelines for Implementation of AI Principles	2021	Asia-Pacific	public sector
Ministry of Economy, Trade and Industry: AI Governance in Japan Ver. 1.1	2021	Asia-Pacific	public sector
NA: Artificial Intelligence in Healthcare	2021	Asia-Pacific	public sector
Government of Vietnam: National Strategy On R&D and Application of Artificial Intelligence	2021	Asia-Pacific	public sector
Government of Australia: Australia's AI Action Plan	2021	Asia-Pacific	public sector
Government of Turkey: NATIONAL ARTIFICIAL INTELLIGENCE STRATEGY	2021	Asia-Pacific	public sector
Caretaker Deputy Prime Minister of Armenia Tigran Avinyan: Artificial Intelligence Strategy for Armenia	2021	Asia-Pacific	public sector
Office of the Privacy Commissioner for Personal Data: Guidance on the Ethical Development and Use of Artificial Intelligence	2021	Asia-Pacific	public sector
Ministry of Science, Technology & Innovation (MOSTI): Malaysia National AI Roadmap 2021-2025	2021	Asia-Pacific	public sector
Capgemini: Our Code of Ethics for AI	2021	Europe	business
EIOPA: Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the european insurance sector	2021	Europe	business

Ada Lovelace Institute: Three proposals to strengthen the EU Artificial Intelligence Act	2021	Europe	civil society
Access Now and others: An EU Artificial Intelligence Act for Fundamental Rights	2021	Europe	civil society
European Digital Rights (EDRi): Open letter: Civil society call for the introduction of red lines in the upcoming European Commission proposal on Artificial Intelligence	2021	Europe	civil society
Privacy International: Privacy International's submission for the UN High Commissioner for Human Rights' report on the right to privacy and artificial intelligence	2021	Europe	civil society
Council of Europe: Guidelines on Facial Recognition	2021	Europe	public sector
European Parliament: Artificial intelligence: questions of interpretation and application of international law	2021	Europe	public sector
GCHQ: Pioneering a New National Security	2021	Europe	public sector
Barcelona City Council: Government measure for a municipal algorithms and data strategy for an ethical promotion of artificial intelligence	2021	Europe	public sector
Government of Scotland: Scotland's AI Strategy	2021	Europe	public sector
Government of Ireland: AI - Here for Good: National Artificial Intelligence Strategy for Ireland	2021	Europe	public sector
UK AI Council: AI Roadmap	2021	Europe	public sector
UK Government: National AI Strategy	2021	Europe	public sector
Government of Italy: Strategic Programme on Artificial Intelligence	2021	Europe	public sector
PwC/WEF: 9 ethical AI principles for organizations to follow	2021	Global	business
EY: Making Artificial Intelligence and Machine Learning trustworthy and ethical	2021	Global	business
Linux Foundation: AI & Data Principles for Trusted AI	2021	Global	civil society
NATO: NATO Artificial Intelligence Strategy	2021	Global	public sector
UNESCO: Recommendation on the Ethics of Artificial Intelligence	2021	Global	public sector
World Health Organization (WHO): Ethics and governance of artificial intelligence for health	2021	Global	public sector
Hewlett Packard Enterprise: HPE AI ETHICS AND PRINCIPLES	2021	North America	business
Adobe: Adobe's Commitment to AI Ethics	2021	North America	business
Salesforce: AI Ethics Maturity Model	2021	North America	business
Meta: Facebook's five pillars of Responsible AI	2021	North America	business
BCG: Responsible AI Builds Trust in Government	2021	North America	business
PwC: Ethical AI: 10 principles the world (mostly) agrees on — and what to do about them	2021	North America	business
Accenture: Responsible AI: From principles to practice	2021	North America	business
AI Now Institute: Algorithmic Accountability for the Public Sector – Report	2021	North America	civil society
OBVIA and Algora Lab: Responsible Artificial Intelligence: a Guide for Deliberation	2021	North America	civil society
Defense Innovation Unit: RESPONSIBLE AI GUIDELINES IN PRACTICE	2021	North America	public sector
Treasury Board of Canada Secretariat: Responsible use of automated decision systems in the federal government	2021	North America	public sector
Government of Peru: National Artificial Intelligence Strategy	2021	South America	public sector
Government of Brazil: Summary of the Brazilian Artificial Intelligence Strategy	2021	South America	public sector
Government of Colombia: ETHICAL FRAMEWORK FOR ARTIFICIAL INTELLIGENCE IN COLOMBIA	2021	South America	public sector
Government of Chile: Política Nacional de Inteligencia Artificial	2021	South America	public sector
SDAIA: Principles of AI Ethics	2022	Africa and Middle East	public sector
SenseTime: AI Governance for Balanced Development	2022	Asia-Pacific	business
GPAI: GPAI 2022 Ministers' Declaration	2022	Asia-Pacific	public sector
Nokia: Nokia's 6 Pillars of Responsible AI	2022	Europe	business
Yandex: Yandex AI Principles	2022	Europe	business
Statworx: statworx AI Principles	2022	Europe	business
RELX: Responsible Artificial Intelligence Principles at RELX	2022	Europe	business

World Economic Forum: A Blueprint for Equity and Inclusion in Artificial Intelligence	2022	Europe	civil society
Future of Life et al.: Call for better protections of people affected at the source of the AI value chain	2022	Europe	civil society
Department for Digital, Culture, Media and Sport: Establishing a pro-innovation approach to regulating AI	2022	Europe	public sector
Government of Belgium: NA	2022	Europe	public sector
The International Federation of Pharmaceutical Manufacturers & Associations (IFPMA): IFPMA Artificial Intelligence Principles	2022	Global	business
UN CEB: Principles for the Ethical Use of Artificial Intelligence in the United Nations System	2022	Global	public sector
BCG: AI Code of Conduct	2022	North America	business
Juniper Networks: AI Innovation Principles	2022	North America	business
Illumina: Ethical Artificial Intelligence Principles	2022	North America	business
ADP: AI Ethics Statement	2022	North America	business
HuggingFace: Putting ethical principles at the core of the research lifecycle	2022	North America	business
Cisco: Cisco Principles for Responsible Artificial Intelligence	2022	North America	business
Government of Ontario: Beta principles for the ethical use of AI and data enhanced technologies in Ontario	2022	North America	public sector
US Government: Blueprint for an AI Bill of Rights	2022	North America	public sector

Table 3.12: *List of AI policy and ethics documents included in the analysis.*

Conclusion

In the summer of 2024, global AI policy remains a fast-evolving subject of study, whose relevance is growing constantly as AI makes inroads into more and more aspects of our lives. In light of recent technological advances and ahead of numerous landmark elections around the world, few would dispute the topic's salience, relevance, and timeliness. Yet, academic scholarship is struggling to keep pace with the breakneck speed of developments. At a time when societies are faced with the challenge of designing the future path of AI, it is paramount to rely on a sound scientific understanding of the myriad socio-economic impacts the technology can bring about – as well as the political and legal frameworks and debates within which policies are being formulated.

This dissertation project aimed to feed into this body of knowledge by providing conceptual, methodological, and empirical contributions. Taken together, its three articles enhance our understanding of the main actors engaged in the nascent field of global AI policy, focusing on how they approach and frame relevant aspects based on their respective preferences. To better understand the underlying approaches and processes, the dissertation offers distinct perspectives – theoretically, methodologically, and substantively – that shine light on various key dynamics and stakeholders. The extensive data collection efforts and subsequent qualitative and quantitative document analyses offer novel empirical evidence to substantiate discussions on global AI policy and governance. In doing so, the papers contribute to various literatures and broaden the knowledge base for scholars and policymakers wishing to better understand important developments in this fast-moving space.

The papers speak strongly to global governance perspectives, which are focused on understanding the intricate web of international stakeholders that collectively shape the governance structures and policies. As became evident throughout the dissertation, global AI policy is currently in the making – against the backdrop of a nascent, but still fragmented global AI governance landscape. Regarding the design of the wider global governance regime, actors can turn either to creating new institutions and mechanisms or accommodating new challenges within existing frameworks. Approaches can differ in the issues that are prioritised or addressed at all. Additionally, the way that a political actor frames a certain issue reveals important information about their preferences, concerns, and priorities. As the articles have shown, different actors prefer and pursue

different approaches to global AI policy, based on their respective interests and distinct roles. These differences are structured both by the specific sector that a stakeholder pertains to, and by geography. Notably, preferences and the corresponding framings are not static, but may change over time. In this regard, the research has shown that the public sectors are most prone to adjustments and that there are signs for increasing alignment between the American and European camp.

Summary of the findings and contributions

Below, I briefly summarise the main findings of each article and situate the results of my research in a broader discussion of AI governance. In doing so, I draw some overarching conclusions for scholarship and outline avenues for future research.

The first article embarked on understanding the nature and trajectory of the nascent global AI governance regime. Initiatives can be structured along two dimensions, (i) state-led vs. non-state-led, and (ii) embedded in existing architecture vs. establishing new instruments. Amidst the fragmented landscape, I found a clear tendency to accommodate governance initiatives within existing structures, both by state and non-state actors. Despite a growth in international agreements and declarations, there are, however, still no legally binding treaties yet. Labour was – and continues to be – distributed fairly equitably between national governments and international organisations. Whereas the latter moved rather early and thus exhibited considerable competence vis-à-vis member states, the former seem to accept their role as useful for international collaborations. Perhaps more recent events suggest a growing inclination by national governments to steer the agenda, as exemplified by the involvement of high-level representatives at the UK's AI Safety Summit. Future research should carefully consider this to update previous findings. Lastly, regarding existing international standards organisations, I found increased attention from key governments such as China, the EU, and the US. Sub-state actors from the public sector were practically not present.

Although the nascent AI regime is fragmented because there is substantial overlap in different actors' membership and the topics addressed by these initiatives, the analysis revealed some degree of centralisation and coordination, with the OECD emerging as a pivotal reference point. I argue that the OECD has considerable epistemic authority and norm-setting power, leading to the first signs of consolidation amongst international actors. Interestingly, the OECD, as a non-state actor, hosts the GPAI – a state-led initiative which can be considered the most advanced AI governance instrument to date. The OECD's role as a host of the GPAI plays a major role in avoiding policy incoherence and further fragmentation. Recent outcomes of the UK's AI Safety Summit suggest that the international AI community intends to continue its collaborative path through the OECD, the GPAI, and beyond.

Future research in this global AI governance literature should turn attention to studying the concrete outcomes of these arrangements. Beyond epistemic and agenda-

setting power, what role does the OECD and the GPAI play in shaping international and domestic AI regulation? How successful are these organisations in fostering impactful global cooperation? Moreover, in light of the Western-centric nature of these two entities, what avenues are there to better integrate the Global South into the debate? And how to deal with the role of China, which is one of the key players in AI technologies but has been following a very different approach, which some fear may lead to a further bifurcation of technological ecosystems? Scholars and policymakers alike will have an interest in answering these questions as the global AI governance regime evolves.

With the second article, I took a closer look at one special actor type, namely national governments. Concretely, I studied in which way governments address AI's impact on democracy. The qualitative in-depth analysis of national AI strategies of 29 OECD member states relied on a novel analytical framework, containing seven dimensions of democracy: participation, equality, civil liberties, vertical accountability, diagonal accountability, responsiveness, and effectiveness. This modified version of Diamond and Morlino (2004)'s conceptual framework provides a useful heuristic tool in quantitative document analysis for topic and stance analysis. On a conceptual level, for each democracy dimension, I presented various potential levers or mechanisms through which democracy may be impacted by the development and roll-out of AI, containing both direct and indirect effects. Additionally, I explicitly state the risks and opportunities associated with each dimension. Applying this framework in a systematic review of the 29 selected national AI strategies, I found that a majority of strategies do not explicitly address many of AI's possible impacts on democracy at all. If stated, governments often remain generic and superficial. However, the study revealed a high variation across the different dimensions and associated framings. One dimension that receives relatively much attention refers to civil liberties, such as privacy issues. Notably, this fundamental dimension is coupled with a very large number of neutral references. The second-most featured dimension is equality. Several countries seem to be concerned about discrimination and equal treatment, but only a few directly link it to democracy and political processes. The most consensual aspect is AI's potential for enhancing the public sector's performance, which is acknowledged in all strategies and framed exclusively in positive terms. In sum, there is a high variation in how extensively government strategies deal with AI's impact on democracy; certain states, such as Spain, Poland, New Zealand or Italy, deal with a broad range of issues, covering multiple dimensions. Contrarily, countries like Australia, Chile, or Lithuania show little engagement. Based on the prevalence of positive and negative framings, countries can be classified into optimistic, sceptic, and ambivalent camps. While the three camps are of similar size, the optimistic group stands out for comprising comparably smaller countries. Lastly, considering the date of publication of the national strategies, I found that earlier strategies are more likely to be positive in their coverage of democracy-relevant issues, whereas most of the recent ones fall into the sceptic or ambivalent camp.

Overall, the article's contributions enrich the academic discourse on the relation-

ship between AI and democracy, providing valuable insights for scholars, policymakers, and practitioners engaged in AI governance and the preservation of democratic principles in the digital age. By presenting and applying a novel analytical framework, it lays the groundwork for further research on the complex interplay between AI technologies and democratic societies. This is a research area that by definition should be multidisciplinary and drawing from a broad range of literatures. The paper hopefully stimulates further discussions on the crucial role of AI in shaping the future of democracy, and the need for societies to pro-actively engage with the various challenges that this poses.

The article also shows the divergence in how seemingly like-minded governments approach important questions of AI governance. It thus confirms previous studies that identified similar differences, e.g., across EU member states (Djeffal, Siewert, and Wurster 2022) or across sectors (Tallberg, Lundgren, and Geith 2023). While fragmentation of approaches around the world may be an obstacle to successful international cooperation, such a competition of ideas and policies may also serve as a useful and efficient way to identify the most robust and reliable responses to the risks and challenges posed by AI, regarding democracy and beyond. As a side note, the analysis exposed an asymmetry in the way that stakeholders think about the harms and benefits of AI. This asymmetry is by no means exclusive to AI's impact on democracy – ambivalent consequences can be expected in many other domains, but are not always adequately addressed in policy discussions.

In the third article, I applied computational text analysis to an original data set of over 300 AI policy documents, to detect topics and stances. The empirical analysis studied sector-specific preferences on one of the most prominent dimensions of political conflict, namely innovation vs. protection, along geographical lines and over time. Notably, in addition to businesses and civil society, I included the public sector as a third actor type, resonating from the stakeholder mapping presented in chapter 1. It showed that in general, across all sentences and documents, the protection topic was relatively more prevalent than the innovation topic. The aggregate results suggest that both business and civil society documents are more attuned to discussions around protection than the public sector, which speaks twice as often about innovation as the other two sectors. In terms of regional differences, it stands out that European actors display a higher innovation affinity than North Americans.

Comparing developments over time, the analysis suggests that civil society in Europe has shifted towards protection, whereas the public sector has moved the opposite way, towards innovation. The opposite is the case in North America, where the relative attention to protection displayed by public sector documents has experienced a significant rise. Simultaneously, civil society actors have become more alert to innovation. In both regions, the business sector is remarkably constant, demonstrating the lowest share of attention to innovation – surprising at first sight, but indicative of actors' underlying motives to signal their attention to ethical concerns. Additional insights into the stances actors express towards AI reveal substantial differences between sectors,

some of which are moderated by a region's temporal trends. Whereas businesses act as AI's cheerleaders, civil society serves a critical watchdog function. The patterns seem to reflect actors' overall preferences for less and more regulation, respectively. The public sector ranges somewhere in between the other two. The increase in positive stance in Europe suggests an embrace of the technology by EU policymakers. Comparing developments within sectors, I find additional evidence of the remarkable shift of the public sector in Europe towards an increased innovation-affinity. In both Europe and North America, civil society has moved towards even more negative stances. All things considered, while there are clear differences between the American and European approaches to regulating AI, recent developments – which contextualise the data – indicate a certain alignment or convergence across the Atlantic.

Synthesising the insights of the three articles, a nuanced understanding of the global AI policy landscape emerges - covering governance structures, policy priorities, and stakeholders' attitudes towards different regulatory approaches. The centralisation around the OECD underscores the importance of leveraging existing governance structures, allowing for adaptive responses to emerging challenges such as the one induced by new technological developments. At the same time, the identified blind spots in national strategies highlight the necessity for critical debate that features inclusive dialogue with citizens and civil society watchdogs, and the introduction of public policies that prioritise the protection of democratic principles and human rights. Putting the insights of the first and second paper together, it is clear that democratic states should cooperate at the global level to promote AI developments aligned with their vision of free, democratic society. However, as the third paper indicates, sectoral and geographical differences in stakeholder preferences on AI regulation show the challenges in finding agreement on how to steer the technology. Promisingly, recent developments hint at avenues for more alignment on key policy and regulatory issues, making room for increased transatlantic cooperation on AI governance.

Overarching elements

Across the three articles of this thesis, there are some recurring elements. One relates to the importance of framing and the setting of expectations in this nascent policy area. The future development of emerging technologies and their respective governance regimes is heavily conditioned by the first stage of the policy cycle, i.e., by processes of agenda-setting and issue definition (Gilardi, Shipan, and Wüest 2021; Schiff 2022; Ulnicane et al. 2022). Within this, there is a strong dual dynamic of competition and convergence. As Jobin, Ienca, and Vayena (2019) pointed out, stakeholders in general seem to agree on the topics that need to be considered in the AI ethics debate. However, they portray varying interpretations and conclusions as to why certain ethics principles are deemed important, or how they should be implemented and by whom (ibid.). My analysis of stakeholder preferences adds additional evidence to this, identifying differences (i.e., competition) but also hinting at alignment (i.e., convergence). Considering both levels, the national and the international, the results thereby suggest that framings and narratives prevailing globally and domestically are mutually reinforcing. We see state-led initiatives bringing their preferences to the international agenda, shaping the governance structures and policy discussions accordingly. Simultaneously, the international convergence in return leads to a certain alignment of nation-state preferences. One example is the OECD, which acts as a reference point setting a direction in the fragmented, polycentred governance structures.

The dissertation also demonstrates the activism exhibited by various stakeholders in recent years when it comes to global AI policy. Mirroring the remarkable pace, breadth, and scope of technological advances, AI policy discussions canvass a broad variety of issues and discuss them from many different perspectives. Still, though, the political fault lines of political conflict often remain obscure, which reflects the nascent stage of political discussions on the topic. In that regard, the dissertation traces the evolution of a tentative structuring of the debate: global developments described in chapter 1 are largely consensual, general, and rather vague in nature. This highlights the shared problem-awareness amongst key actors, but also makes clear the lack of concrete policy as potential winners and losers of different choices were not readily identifiable. In turn, chapter 2 already identifies considerable variation across national AI strategies regarding one specific aspect of AI (democracy), which hints at broader differences in national approaches. Inter alia, democratic governments do not seem to share a common problem awareness – though there seems to be consensus about the potential benefits of AI for improving public sector efficiency. Lastly, the third chapter showed the challenges in revealing fault lines: comparing innovation- vs protection-attuned positions for the two sides of the Atlantic resulted in a largely kaleidoscopic panorama with sometimes unexpected and detrimental trends. Arguably, stakeholders are still going through the process of identifying and updating their preferences, formulating the corresponding policy positions, and following through on them through public policy actions.

Efforts at identifying lines of political conflicts are confronted with some fundamental conceptual challenges all too common in political science research. To explain and theorise observed phenomena requires the construction of categories, which in the messy reality of the social world too often turns into an arbitrary or at least fuzzy enterprise. For instance, I considered regulatory sandboxes as an example to illustrate an innovation-friendly approach, since these are legal instruments that grant companies and developers a relatively high degree of freedom to live-test their systems. However, as these sandboxes are usually restricted and supervised, one might plausibly consider them examples of a protection-leaning approach, aimed to prevent unguarded deployments. Ultimately, the classification depends on the context and the particularities that govern a specific sandbox regime, as well as the criteria established by the researcher for defining either approach.[18]

As the global discourse on AI governance intensifies, the dissertation's findings contribute not only to understanding the complex interplay of actors, their preferences and framings but also to guiding future research endeavours. AI is a fast-developing technology and as this dissertation has demonstrated, the accompanying legal, political, and regulatory responses are only beginning to shape up. As such, it is hardly surprising that most current research on AI governance and policy leans towards exploratory, descriptive, or prescriptive perspectives. This includes my dissertation. Throughout it, I have stressed that findings should be seen as indicative, not conclusive. Accordingly, there is ample room for future research to build on the work presented here. This includes improvements to the data collection and the methods. The computational text analysis undertaken in chapter 3 has shown the promises as well as challenges associated with using LLMs for social science research. While such models should be addressed with caution and a healthy dose of scepticism, the technology and its applications will continue to improve rapidly, opening doors to an exciting and highly promising addition to the researcher's toolkit. Throughout this process, the field should pay attention to developing and adhering to a comprehensive toolkit which minimises potential risks. Chapter 3 includes elements of such a toolkit (e.g., repeating the LLM's prediction task multiple times and then taking averages rather than relying on individual output; benchmarking outputs against samples labelled by humans). In these efforts, the growing AI policy literature is not inherently different from other social science fields and should follow similar methodological considerations.

Over time, researchers should also increasingly focus on different types of questions that shift the focus from descriptive to more analytical investigations. Beyond the various research avenues mentioned in the three articles, there is an urgent need to dig deeper into policymaking processes and to study factors that may be less directly observable than documents and framing. There is justified concern that many actors,

[18] The author is deeply grateful to the two thesis reviewers for their insightful and invaluable feedback that highlighted the permanent reflexivity this research warrants.

especially from the business world, engage in ethics washing or meaningless virtue signalling. Critical research needs to interrogate such vanity efforts. For that, it will be essential to tap into novel data sources that provide more accurate and complete information into competing interests, actions, and outcomes. Following Veale, Matus, and Gorwa (2023)'s "cui bono?" question, critical literature should investigate the net beneficiaries of regulatory (in-)action, together with their motivations and rationales. The work by Schiff, Laas, et al. (2022) can serve as a blueprint in this regard. Critically, the attention to causes and consequences beyond framing must consider plans and measures taken by actors away from the limelight. When do discrepancies between lofty discussions and concrete governance dilute meaningful outcomes? Do AI models reflect the letter and spirit of international agreements, recommendations, or other publicly issued commitments? Are laws and treaties enforceable and do they get enforced?

Asking and answering these questions will inevitably raise pertinent normative discussions, which academic scholarship should not refrain from. The remaining paragraphs offer the reader a glimpse into the author's perspectives regarding some of the key debates: More often than not, it seems as if technological developments shape the policy discourse, and not the other way around. Societies should seriously consider whether it ought to be like that – and react accordingly. For instance, the lengthy and thorough procedure of developing and negotiating the EU's AI Law was completely overtaken by developments of LLMs and generative AI (Hacker, Engel, and Mauer 2023). Proponents of technological advances often paint them as inevitable and denounce rigid regulations as innovation-hindering. But in light of the profound and largely unforeseeable consequences that some innovations are bound to bring upon the world, it may be wise to move slowly and cautiously, occasionally pausing to allow for collective reflection, sufficient scrutiny, prudent foresight, critical debate, and democratically legitimised decisions.

In addition to more rigorous causal research, scholars should continue to interrogate the fundamental concepts of global governance and politics in light of the transformations brought about by AI technologies. For instance, Erman and Furendal (2022) discuss implications of AI on the political legitimacy of global governance, arguing that research should take into account procedural aspects of good governance just as much as it should study the effects and outcomes of governance mechanisms. One instance of this that I mentioned earlier might be the integration of civil society and ordinary citizens into AI policy processes (c.f. Schiff, Borenstein, et al. 2021; Stix 2021). Some countries and organisations have taken promising steps in this regard. For instance, the Chilean strategy was subject to extensive and meaningful participatory campaigns and consultations. And the inclusive deliberation process that led to the landmark "Montreal Declaration for a Responsible Development of Artificial Intelligence" illustrates how productive dialogue between citizens, experts, public officials, industry stakeholders, civil organisations, and professional associations can be. However, too much of global AI governance is still happening behind closed doors or far away from the public

(for an analysis of the US case, see Schiff 2023). The recent attention to AI will likely increase the pressure on stakeholders to increase transparency and lead to more accountable and inclusive processes, which future research should follow closely.

Lastly, Tallberg, Erman, et al. (2023) outline a useful research agenda that distinguishes between empirical and normative research. They emphasise the importance of studying regulatory issues as well as the underlying interests, processes, and consequences. One of the key challenges for the field moving forward will then be to productively and insightfully combine those different perspectives. Multidisciplinary mixed-methods research and extensive collaboration across geographic and scholarly borders will be essential in achieving such a truly global and holistic understanding of how global AI policy is made (see also Büthe et al. (2022) who sketch out an interdisciplinary research agenda in more detail).

Coincidentally, leading research labs developing cutting-edge AI technologies feature exactly these characteristics. It therefore seems only logical that scholars from political science, international relations, legal and technology studies, or other interested researchers should follow this approach when trying to describe, measure, analyse, and explain the multifaceted and fast-evolving political, legal, ethical, and regulatory developments around global AI policy. And perhaps use a bit of AI to support their research, where pertinent.

Bibliography

Büthe, Tim et al. (2022). "Governing AI – attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence". In: *Journal of European Public Policy* 29.11, pp. 1721–1752. DOI: 10.1080/13501763.2022.2126515 (cit. on p. 147).

Diamond, Larry Jay and Leonardo Morlino (2004). "The Quality of Democracy: An Overview". In: *Journal of Democracy* 15.4, pp. 20–31. DOI: 10.1353/jod.2004.0060 (cit. on p. 141).

Djeffal, Christian, Markus B. Siewert, and Stefan Wurster (2022). "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies". In: *Journal of European Public Policy* 29.11, pp. 1799–1821. DOI: 10.1080/13501763.2022.2094987 (cit. on p. 142).

Erman, Eva and Markus Furendal (2022). "Artificial Intelligence and the Political Legitimacy of Global Governance". In: *Political Studies* (cit. on p. 146).

Gilardi, Fabrizio, Charles R. Shipan, and Bruno Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". In: *American Journal of Political Science* 65.1, pp. 21–35. DOI: 10.1111/ajps.12521 (cit. on p. 144).

Hacker, Philipp, Andreas Engel, and Marco Mauer (2023). "Regulating ChatGPT and other Large Generative AI Models". In: *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123. DOI: 10.1145/3593013.3594067 (cit. on p. 146).

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9. ZSCC: 0000007, pp. 389–399. DOI: 10.1038/s42256-019-0088-2 (cit. on p. 144).

Schiff, Daniel (2022). "Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy". PhD thesis. DOI: 10.17605/OSF.IO/KW8XD (cit. on p. 144).

— (2023). "Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy". In: *Review of Policy Research* 40.5, pp. 729–756. DOI: 10.1111/ropr.12535 (cit. on p. 147).

Schiff, Daniel, Jason Borenstein, et al. (2021). "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection". In: *IEEE Transactions*

on Technology and Society 2.1, pp. 31–42. DOI: 10.1109/TTS.2021.3052127 (cit. on p. 146).

Schiff, Daniel, Kelly Laas, et al. (2022). "Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies". In: *Codes of Ethics and Ethical Guidelines*. Ed. by Kelly Laas, Michael Davis, and Elisabeth Hildt. Vol. 23. Cham: Springer International Publishing, pp. 121–143 (cit. on p. 146).

Stix, Charlotte (2021). "Foundations for the future: institution building for the purpose of artificial intelligence governance". In: *AI and Ethics*. DOI: 10.1007/s43681-021-00093-w (cit. on p. 146).

Tallberg, Jonas, Eva Erman, et al. (2023). "The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research". In: *International Studies Review* 25.3. DOI: doi.org/10.1093/isr/viad040 (cit. on p. 147).

Tallberg, Jonas, Magnus Lundgren, and Johannes Geith (2023). *AI Regulation in the European Union: Examining Non-State Actor Preferences*. arXiv:2305.11523 [econ, q-fin]. DOI: 10.48550/arXiv.2305.11523 (cit. on p. 142).

Ulnicane, Inga et al. (2022). "Governance of Artificial Intelligence". In: *The Global Politics of Artificial Intelligence*. 1st ed. Boca Raton: Chapman and Hall/CRC, pp. 29–56. DOI: 10.1201/9780429446726-2 (cit. on p. 144).

Veale, Michael, Kira Matus, and Robert Gorwa (2023). *AI and Global Governance: Modalities, Rationales, Tensions*. DOI: 10.31235/osf.io/ubxgk (cit. on p. 146).