Silicon Valley and the Moral Landscape of Artificial Intelligence Governance in the European Union and the United States

A Dissertation submitted by

Sarah Hladikova

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Interdisciplinary Studies

TUFTS UNIVERSITY

Primary Advisor: Professor Peter Levine

Advisory Committee: Professor Lenore Cowen, Professor John Shattuck

Silicon Valley and the Moral Landscape of Artificial Intelligence Governance in the European Union and the United States

Abstract

Artificial intelligence (AI) as a socio-technical phenomenon has become a prevalent subject of scientific research across disciplines. AI has also become an internationally recognized governance challenge and a source of ongoing fascination across mainstream media and popular culture. In Western democracies, the lack of democratic oversight of AI systems has become a prominent concern in the criticism aimed at the private sector for its dominant position and dubious self-governing efforts. At the same time, governments' interventions were doubted for their slow progress and perceived lack of expertise.

Since the early 2010s, technology companies' entrenched position in AI development and deployment across societal domains and functions has presented a problem for regulators and civil society who wish to provide input and participate in designing domain-specific or general-purpose AI applications.

This dissertation addresses current tensions surrounding AI governance through quantitative and qualitative analyses. First, I analyze existing AI policy initiatives, providing a comprehensive understanding of the moral landscape of AI governance. Second, this thesis presents a compelling case study that bridges the gap between AI practitioners and those involved in AI ethics and governance. By

integrating interdisciplinary research methods and insights from practitioners, this dissertation contributes to developing the AI governance field and refining its methodology and practices.

Acknowledgments

One of my professors at Tufts University, Daniel Dennett, once wrote that "if you can approach the world's complexities, both its glories and its horrors, with an attitude of humble curiosity, acknowledging that however deeply you have seen, you have only scratched the surface, you will find worlds within worlds, beauties you could not heretofore imagine, and your own mundane preoccupations will shrink to proper size, not all that important in the greater scheme of things" (Dennett, 2007).

I am so lucky to have many incredible people around me to thank for making the years in graduate school some of the most inspiring, motivating, and intellectually stimulating of my life. Their incredibly diverse thoughts and interests sparked my curiosity, and their brilliance kept me humble.

First and foremost, I am incredibly thankful to my primary advisor and advisory committee. Your thoughtful advice and trust in my abilities motivated me and gave me the confidence to design this project in a way I see most meaningful. Thank you to my advisor, Peter Levine, for giving me the freedom to carve my own path and supporting me in moments when I inevitably hit some brick walls. Having you in my corner made all the difference, and I am forever grateful for that. Thank you to my advisory committee members, Lenore Cowen and John Shattuck, who were immensely generous with their time and attention over the years. Your support, insights, and thoughtful advice helped me see the topic I care about deeply from different angles. I learned so much from you.

Thank you to my mentors, Andreia Martinho, Parnian Mokri, Susan Landau, and Natasha Warikoo, for all your wisdom, kindness, and support. You were my source of inspiration and empowerment. I am also grateful to all my professors and colleagues, my research networks, and groups at Tufts University and beyond for many enriching and fascinating conversations. Thanks to you, I learned so much and became a better person.

Thank you to my family, my mom, Zuzana Hladíková, for always believing in me and for your unconditional love. Thank you to my father, Vladimír Hladík, and my sister, Carolina Hladíková, for their patience and support. Thank you to my grandfather, Vladimír Hladík Sr., for raising me to value education, democracy, and freedom above all else.

Finally, thank you to my fiancé, Nicholas Krebs. I would not be writing this if it wasn't for you. You made the most challenging years of my life the happiest.

Chapter 4 of this dissertation has been previously published as:

Hladikova, S., Wang, Y., & Martinho, A. (2024). The Third Moment of AI Ethics: Developing Relatable and Contextualized Tools. Manuscript under review.

Hladikova, S., Wang, Y., & Martinho, A. (2024). Integrating Ethical Reflection in AI Development: Challenges and Lessons from Developing an AI Ethics Tool. In: Envisioning Ethics, EASST-4S 2024, Amsterdam.

I thank my co-authors for their kind permission to include this work in my dissertation.

Dedication

To Nicholas.

And for Lucky.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

"*Technology—that child of modern science, which in turn is a child of modern metaphysics—is out of humanity's control, has ceased to serve us, has enslaved us and compelled us to participate in the preparation of our own destruction. And humanity can find no way out: we have no idea and no faith, and even less do we have a political conception to help us bring things back under human control.*"

- Vaclav Havel, The Power of the Powerless

Artificial intelligence (AI) is a powerful and pervasive technology that has the potential to transform various aspects of human life. However, civil society is largely excluded from the governance of AI[1], leaving important design and deployment decisions to expert groups, corporations, and (sometimes) governments. This situation is exacerbated by the negative portrayal of AI in popular media, academic research, and even significant parts of industry, which often emphasize the risks and dangers of AI rather than the opportunities and benefits. The debate over the nature of AI risks is complex and diverse, ranging from long-term, speculative existential threats to humanity (Bostrom, 2014;

[1] The term governance does not have one agreed upon definition. The United Nations' definition of governance is "the exercise of economic, political and administrative authority to manage a country's affairs at all levels. It comprises the mechanisms, processes and institutions through which citizens and groups articulate their interests, exercise their legal rights, meet their obligations and mediate their differences" (The United Nations Development Programme, 1997). Mark Bevir writes: "Governance refers, therefore, to all processes of governing, whether undertaken by a government, market, or network, whether over a family, tribe, formal or informal organization, or territory, and whether through laws, norms, power or language. Governance differs from government in that it focuses less on the state and its institutions and more on social practices and activities" (Bevir, 2012). Francis Fukuyama defines governance as a "government's ability to make and enforce rules, and to deliver services, regardless of whether that government is democratic or not." (Fukuyama, 2013).

Consequentially, AI governance does not have one agreed upon definition either. Adopting Joanna Bryson's definition of AI as "any artefact that extends our own capacities to perceive and act" (Bryson, 2019), and starting from the above definitions, I define AI governance as **the laws, norms, power or language that articulate stakeholders' interests, exercise their legal rights, meet their obligations and mediate their differences over development, deployment, and use of artefacts that extend human capacities to perceive and act.**

Russell 2019; Ord 2020) to more immediate ethical and social challenges such as bias, discrimination, and environmental impact (Buolamwini, 2023; Buolamwini & Gebru 2018; Gebru, 2020; Bender et al., 2021).

The current debates on AI risks, however, are not sufficient to enable civil society to participate in the governance of AI. Rather than trying to understand civil society's preferences and values regarding AI, these debates tend to be dominated by experts and elites, often with vested interests. As a result, elected officials and policymakers tend to adopt a risk-averse approach that prioritizes the prevention of catastrophic outcomes over the promotion of positive ones. This approach often focuses on the most advanced or hypothetical forms of AI, such as *artificial general intelligence* (AGI) or *superintelligence*, which may neglect the more immediate and tangible impacts of AI on various domains of society (Galanos, 2018; Schopmans, 2022). Moreover, this approach seems to be often supported by industry leaders and industry-backed think tanks, who may have an incentive to influence the regulatory agenda in their favor. This raises a question of regulatory capture, which occurs when a regulatory agency serves the interests of the regulated industry rather than the public interest (Carpenter & Moss, 2013; Toner & Fist, 2023; Dal Bo, 2006).

This dissertation explores the role of non-state actors, particularly international corporations, in shaping the moral landscape of AI and its governance. In this work, I focus on international corporations for three reasons. First, the focus on tech companies allows for a comparative perspective of the US and the EU markets, as the tech sector has highly concentrated power, with a few

very dominant actors offering their services to billions of users worldwide.[2] Second, studying international corporations allows us to examine the relationship between self-governance and government intervention (Sørensen & Triantafillou, 2016). Finally, among the non-state actors in AI governance, international corporations have been the most vocal proponents of specific approaches and the most prominent subject of state and international regulatory efforts (Schiff et al., 2020).

Expanding upon Abend's (2014) definition of a *moral background*, I refer to the implicit assumptions and values that shape ethical debates and decisions about AI and investigate how, notably, the themes of ethics and risk emerged and influenced the governing efforts in the AI field.

I study the moral landscape from an interdisciplinary perspective. As Artificial Intelligence (AI) continues to advance, its impact on society grows exponentially. From autonomous vehicles to lethal weapon systems, AI enters various domains, offering economic efficiency, everyday convenience, and a vast amount of accessible information in the blink of an eye. However, alongside these benefits come concerns of misuse and harm to individuals and societies. Hence, the governance of AI demands attention. While there is considerable literature on AI, the specific challenges of AI governance remain underdeveloped. Interdisciplinary research is essential for navigating this complex environment, i. By integrating insights from diverse fields such as law, ethics, computer science,

[2] According to statista, "Google, Amazon, Meta (formerly known as Facebook), Apple, and Microsoft, [f]ormerly known under the acronym GAFAM, now GAMAM or GAMMA, the five tech giants boast user bases in the billions and a combined market value of almost seven trillion U.S. dollars, making them the largest internet companies worldwide" (Clement, 2024).

and public policy, we can better understand the socio-technical challenges inherent to this technology (Taeihagh, 2021).

I quantitatively analyze three corpora of AI-related policy initiatives by the European Union, the United States, and the private sector to illustrate how these actors constructed and contested the moral landscape of AI. The quantitative analysis provides valuable insights into some of the most dominant themes within each of the corpora over the period from the early 2010s until the early 2020s. Then, I conduct a qualitative analysis of the two most dominant themes of the moral landscape; one encompassing ethical concerns in AI, and the second one includes the more speculative, catastrophic concerns that often dominate the AI narrative in popular culture and mainstream media.

This mixed-methods research aims to reflect on the power dynamics that influence the governance of AI and to demonstrate the gap between democratic ideals and the practices of AI regulation. I compare two different regulatory contexts, the European Union and the United States, to investigate how the moral landscape of AI varies across these regions and how it is reflected in their practical approaches to AI governance. Finally, I propose a reflection on the AI ethics community and the ongoing tensions between AI ethics researchers and AI practitioners and offer a new perspective on the use of AI ethics research by practitioners.

The implications of AI's moral landscape are then illustrated by a specific AI application: large language models. These powerful AI systems that can generate natural language and other types of content have raised various ethical

concerns. The final chapter examines how AI's moral landscape shapes the governance efforts of large language models in the EU and the US.

Literature Review and Rationale

AI & Industry Self-Governance

Artificial intelligence (AI) has become associated with the accelerating technological progress of the digital age. The hopeful vision of the late 1950s (McCarthy, 1955) that intelligent computers bring insight into where the human mind cannot reach slowly turned into the sobering reality where technology is either used as a quick fix to societal problems– or with no regard to its impact on society in the first place (Eubanks, 2018).

In Western democracies, the power to make changes and address societal problems traditionally belonged to the government, civil society, and markets. The public interest in the emerging field of computer science research was initially reflected by government funding, and both the American government and its European counterparts supported the establishment of AI research and development. The field, however, encountered significant drawbacks in the 1970s, which led to the withdrawal of government funding. This so-called *AI winter* marks the period of private companies stepping in and taking a leading role in research and development that has remained unchallenged (Radu, 2021).

The booming success that came decades later with the revival of neural networks and the expansion of machine learning across domains meant that tech companies were positioned favorably in the eyes of the public. *Technological determinism*,[3] the notion that technological invention is driven by an unstoppable

[3] Agency is one of the focal points of science and technology studies (Dafoe, 2015). The question we ask is: Who is in charge? "Hannah Arendt (1958: 144) wrote, "[t]ools and instruments

momentum and has the power to reshape society and its values, has become strongly embedded in the public discourse (Jasanoff, 2016). The story of inevitable technological progress has become mainstream with artificial intelligence. Citizens of Western democracies became mere spectators of their future unfolding in front of them, enabled by the benevolent technologists. As a result, the role of governments in shaping the direction and impact of technology, including AI, has been called into question. The debate about the appropriate balance of power between state and non-state actors began to tilt in favor of the latter.

Since tech companies established themselves particularly ahead of states in the research and development of AI, it was not surprising (even welcomed) when these actors started setting their ethical standards and publishing other self-regulatory efforts. The expertise of tech companies provided a new source of legitimacy, and arguments for *technocracy* became prominent in policy and scientific communities. Advanced knowledge in computer science has become a determining factor in one's ability to set standards for cutting-edge AI technology (Jasanoff, 2016). For example, American Senators of the Judiciary and Commerce Committees did not seem fit to govern Facebook after the notorious hearing of

---

are so intensely worldly objects that we can classify whole civilizations using them as criteria." Not only can we, but frequently we do; thus, we speak of the "stone," "iron," "steam," and "computer" ages." wrote Sally Wyatt (2008). Technological determinism implies that technological advancements are inevitable. Scholars of technological determinism note that technology is the defining element of civilizations, and some even argue that its development follows an internal logic that is beyond human control (MacKenzie and Wajcman, [1985] 1999). Other scholars attempt to highlight the complex power dynamics and roles of different actors (Marx & Smith, 1994). Others yet refer to technological determinism as a critics term (Dafoe, 2015).

Mark Zuckerberg in 2018, though their presumably naïve questions might have represented precisely what their constituency wanted to know.[4]

If states rely solely on the self-governance of the private sector, the decisions and practices of a few tech companies will shape the governance of AI writ large.[5] For instance, the self-imposed standards of private companies could satisfy civil society and regulators, and governments may codify private standards into law. Historical evidence from other sectors indicates that industry standards and best practices developed by industry groups, professional organizations, or other non-state actors helped to ensure that the development and use of systems and technologies were aligned with the values and interests of society or that they served well as a complement to state-led approaches to state governance (Maurer, 2017; Josselin & Wallace, 2001).

Industry self-regulation could help ensure that AI systems are developed and used in a way that aligns with society's values and interests if achieving this aligns with their economic incentives. However, self-regulation may not suffice to

[4] "Why am I suddenly seeing chocolate ads all over Facebook? Is Facebook spying on the emails I send via WhatsApp?" were just a few of many confusing questions that Senators asked Zuckerberg during the hearing in the Hart Senate Office Building on Capitol Hill April 10, 2018, in Washington, DC. (Stewart, 2018).

[5] The rationale for self-governance that has mostly impacted the mindsets of current tech leaders in western democracies has been pioneered by commercial communities of the late 19th century (Maurer, 2017). Over the 1980s and 1990s, globalization opened new supply chains and connected markets around the globe. Initially, it was the threat of state intervention that motivated the establishment of private standards, but in global markets, the effort to maintain product interoperability became another early motivator for industry leaders to ensure the markets follow some principal rules (Maurer, 2017). Non-state actors in the tech domain have adopted similar strategies to their predecessors in the early decades of globalization. Their position, however, has become significantly more impactful on international relations. This is both for the reasons of the economic dominance of tech companies (Arnold, Rahkovsky, & Huang; 2020), and because of the technological advancements that they have achieved with artificial intelligence.

address AI's potential risks and impacts, especially in cases where the technology is used in, e.g., healthcare or public safety.

With a lack of democratic oversight or opportunity for deliberation, self-regulatory efforts may further empower tech companies to dictate how the technology is designed and deployed. This position may significantly affect the contestability of established companies, prevent equal access of different advocacy groups to policymakers, or lead to regulatory capture (Carpenter & Moss, 2013; Bietti, 2020; Stigler, 1971). Moreover, artificial intelligence was identified by international relations scholars as a *strategic technology* or *transformative technology* (Ding, 2021; Leung, 2019[6]). The potential impact on the future of democratic societies– both within and about the power structure of international relations writ large– hence necessitates more attention to the tension between non-state actors' self-regulation and government intervention. The detachment between all affected and the few involved in AI governance is striking for technology with such vast potential.

AI & Non-state Actors

Non-state actors are defined by their partial or full autonomy from central government funding and control. These entities often arise from civil society or the market economy and operate or participate across the borders of national states where they are engaged in states' socio-political systems and economies.

[6] Transformative technology refers to a technological advancement that significantly changes society in the way people live or interact with each other. Examples of transformative technology include the internet, smartphones, or electricity (Gruetzemacher & Whittlestone, 2019; Acemoglu & Lensman, 2023).

Their presence directly or indirectly affects the political outcomes of the international system (Josselin & Wallace, 2001).

Non-state actors engaging in the governance of artificial intelligence are primarily international corporations, driven by their economic goals, and expert groups, typically organized within research organizations or think tanks and motivated by shared values and research opportunities. These expert groups, also called epistemic communities, play a significant role in shaping the direction and impact of AI through their expertise and advocacy efforts (Palladino, 2020). There is often tension between international corporations and experts, with the latter group presenting research findings that challenge the former group to improve their practices. This tension is sometimes characterized by an adversarial dialogue between the two groups.

However, the role of experts is not solely in opposition to international corporations; there is also a dynamic of collaboration and exchange of ideas between these actors. Expert groups in socio-technical research areas, such as *AI governance,* may engage in dialogue to identify the most pressing challenges and opportunities related to AI. Some experts within these groups advocate for increased government oversight of AI. Others view state intervention as a potential threat to the scientific advancement of AI, which they argue would essentially mean delaying or preventing the vast benefits this technology may bring to humanity (Prunkl & Whittlestone, 2020; Cave & ÓhÉigeartaigh, 2019).

Moreover, experts in the AI field may hold different views on the research agenda and the priorities for addressing challenges and opportunities related to

AI. Some experts believe that the development of strong artificial intelligence (or artificial general intelligence) poses an existential threat to humanity. Others argue that these concerns are unfounded and that more immediate-impact research needs should be prioritized (Prunkl & Whittlestone, 2020). The influence and role of expert groups in AI governance are uncertain due to many influential actors representing clashing ideas within these groups.

Understanding the role of non-state actors in the governance of AI can be aided by examining the self-governing activities of international corporations, which play a significant role due to their dominant position in research and development (Arnold, et al., 2020). These corporations are driven by economic goals and can exert a major influence on the direction and impact of AI through their resources, expertise, and advocacy efforts. In the US-EU context, the concentration of economic and political power is around a handful of US-based corporations often referred to as *Big Tech*. The group does not have a definite composition; however, it has been understood since the early 2010s as a reference to Google, Apple, Facebook, Amazon, and Microsoft (Kak & Myers West, 2023).

The following criteria were established to identify a set of non-state actors for this research project: a) companies that explicitly contribute to the development of cutting-edge AI-powered technologies by establishing AI research labs, hiring and retaining top talent in AI, and publicly presenting their aim in the AI domain, b) companies that have demonstrated the capability to become a global leader in AI by launching or publishing cutting-edge AI tools or services, c) companies that have established their entrenched and durable position

in the markets by reaching a significant financial evaluation as well as large user base, and d) companies that provide their services to active end users established or located in the EU and the US. Based on the established criteria, the following companies were set as the subject of the analysis; Alphabet Inc. (Google), Amazon.com Inc. (Amazon), Apple Inc. (Apple), Meta Platforms Inc. (Facebook), Microsoft Corporation (Microsoft) and their investee OpenAI. The list overlaps with the collective understanding of Big Tech and the targets of several regulatory efforts, namely the Digital Markets Act in the European Union (European Commission, 2023).

The Moral Landscape of AI Governance

I propose to study the role of international corporations as reflected in the moral landscape of AI governance. The moral landscape encompasses the norms and normative theories invoked in the context of AI (such as beneficence, i.e., "AI should benefit humanity." Or consequentialist theories, i.e., "Consider and manage the possible risks and biases that data sets and algorithms are susceptible to, and how they might affect the outcomes or have unintended consequences" (Hanson et al., 2023).) The moral landscape also includes methods and arguments and the aspects of AI that we subject to moral evaluation (for instance, the accuracy and explainability of a model or even the size of a model became subjects of ethical inquiry).

The moral landscape of AI Governance is motivated by moral background, a term coined by Gabriel Abend (2014). Abend argues that

first-order morality (i.e.: “Lethal autonomous weapon systems (LAWS) should be banned because they are unethical.”) is always underlain by a second-order moral background (I.e.: “LAWS should be banned for use in combat because they cannot be held responsible for their actions.”). While the moral landscape of AI builds upon Abend’s moral background, I intentionally do not use the term “background” for its connotations of being behind the main objects or without a need for input or close attention.[7] I argue that the moral landscape of AI requires close attention and scrutiny, which should be reflected carefully in the language we use when we refer to it.

The study of the moral landscape can inform an investigation into the impact of AI technologists on the distribution of power in society (Greene et. al., 2019). Understanding the moral landscape of AI governance can provide important insights into the limitations of current policy conversations and challenge the mainstream approach to AI ethics of international corporations. The analysis of the moral landscape may also help revise dominant intuitions around AI ethics because, while the moral landscape consists of complex moral, religious, and metaphysical reasonings, they often manifest themselves as intuitive (Abend, 2014). Finally, utilizing the moral landscape framework allows us to reflect on *tech exceptionalism*, the idea that the level of sophistication in technology, particularly AI, is so advanced that it requires a revision of existing

[7] One of the challenges of interdisciplinary research is a careful choice of terms that translate well across disciplines (Bracken & Oughton, 2006). For instance, *background processes* in computation refer to such programs and applications that do not require direct input from the user, such as system monitoring or user notification in a computer operating system.

frameworks or even a new legal and regulatory approach (Allensworth, 2020; Doctorow, 2023; Jones, 2018).

The examination of Abend's moral background is separated into six dimensions: *grounding; conceptual repertoire; object of evaluation; method and argument; metaethical objectivity; and metaphysics*. The *grounding*, according to Abend, is not dissimilar to a philosophical normative theory. It may not be as fully developed to considerably influence a large group of individuals. Abend notes that politicians or news editors can provide an influential grounding without invoking elaborate philosophical arguments. For example, the statement "good is what maximizes my pleasure and/or minimizes my suffering" is as valid grounding as "good is what God commands" is as valid grounding as "good is what makes America great again."

The second dimension of Abend's moral background is *conceptual repertoires*, which allow social actors to recognize the distinction between moral and non-moral. Social actors have access to a repertoire of moral concepts such as purity, exploitation, or appropriateness, allowing them to capture a social phenomenon and assign it a moral value. Different societies have different conceptual repertoires, and they also change over time.

The third dimension is the *object of evaluation*, which reflects the ability to be morally evaluated. The study of this dimension includes investigating what objects are capable and incapable of being morally evaluated, and among these objects of moral evaluation, which ones are evaluated more often, when, where, by whom, and for what purpose.

Method and argument are other dimensions described in the moral background. The study of this layer focuses on the acceptable evidence that one can present to support moral claims and the methods they use to arrive at that conclusion. Evidence for some cultures may be based on scientific discovery, while others may refer to their intuition or gut instinct. The method may be a deduction in some contexts or an analogy in another; the base of those methods may be introspective, spiritual, empirical, and scientific.

The last two dimensions of Abend's moral background are *metaethics* and *metaphysics*. Metaethical theories in philosophy study the nature of ethical statements and their relationship to truth, reality, and objectivity. Some theorists of metaethics argue for moral realism, which asserts that moral statements can be objectively true. Moral realists would say that just as we have historical facts, or empirical observations of the natural world, which we deem to be either factually and objectively correct or incorrect, the same is true for moral facts ("The second world war ended 1945" = "Stealing is always wrong"). Others, moral emotivists, for instance, would disagree with this notion and claim that moral statements are mere expressions of one's subjective preferences ("Oranges are better than mangoes!" = "Good person never lies"). Philosophers study metaethics to establish the real status of moral claims. However, for Abend, the question is which metaethical views people tend to hold. We want to understand whether moral actors believe that their moral judgments are objective or not. The goal of understanding the metaethics in the moral background is not to understand the relationship between moral claims and objectivity, it is to understand whether

moral actors believe their moral judgments are objective. Metaethical assumptions are built into rules and routines, and empirical study can reveal a tacit rejection or endorsement of these assumptions.

Metaphysics is the last dimension of moral background. While the questions he asks reminisce of the questions that philosophers try to answer with their metaphysical theories of the nature of being and reality, his focus is strictly on social metaphysics: "These are the metaphysical pictures or assumptions that ordinary people and social practices, institutions, and understandings manifest. [...] Systems of practices, institutions, and understandings are underlain by metaphysical elements, even though they can be wholly tacit, built into practices, routines, and devices" (Abend, 2014).

Theoretical Framework

This dissertation contributes to the growing interdisciplinary research field of AI governance. It combines methodologies and domain-specific knowledge from computer science and international relations as an interdisciplinary project while adopting a social-scientific and philosophical lens of science and technology studies (Silvast & Virtanen, 2023; Gad & Ribes, 2014).

The emergence of the moral landscape in the context of AI governance can be best understood through Sheila Jasanoff's notion of co-production as a simultaneous process through which modern societies form their epistemic and normative understandings of the world (Jasanoff, 2004; Jasanoff, 2020). In other words, the moral landscape emerged simultaneously with the invention of the

term AI, grew with the first instances of AI as a practice in computer science, and reflects AI as a socio-technical phenomenon I discuss in this thesis. Jasanoff asks, "How is knowledge taken up in societies, and how does it affect people's collective and individual identities, permitting some to be experts, others to be research subjects, and still others to be resistors or revolutionaries?" (Jasanoff, 2004). The co-production framework, coined by scholars in Science and Technology Studies (STS), argues that technological and societal progress are not two separate domains but an intertwined process. The theory of co-production is influenced by Bruno Latour's actor-network theory, which argues that human actors and inanimate objects are all part of a complex network and can influence each other equally. Just as much as we can impact a piece of technology, that piece of technology has an agency to affect us (Latour, 2005). Therefore, the co-production theory does not center humans on technological development, it describes the system in which science, technology, and society coexist and evolve in interrelated networks. The actor-network theory is hence more a method than a theory, as Latour puts it, as it allows one to study a specific phenomenon in its complex evolution over time and concerning the surrounding actors in the network (Latour, 1999).

The co-production perspective in this dissertation means examining how AI governance is shaped by various actors, artifacts, and other factors. The theoretical framework will reflect on the moral landscape of AI in the context of important themes within published policy efforts in the context of technological development. While I argue that technological progress provides an important

nuance to the public discourse and policy debates around AI, I challenge technological determinism and the idea that technology itself defines a certain time period (or the society of that period). According to Latour, agency is not a given quality, but it is the ability to modify other actors through the course of action, and it is hence not limited to humans, as it does not couple agency with intentionality or free will (Latour, 2004). I argue there are levels to agency. Simply put, human agency, as opposed to the agency of an artifact, is motivated by its intentionality, or, as Wendt puts it: "Human agents and social structures are, in one way or another, theoretically interdependent or mutually implicating entities. [...] It is then a plausible step to believe that the properties of agents and those of social structures are both relevant to explanations of social behavior" (Wendt, 1987). An agency of individual nodes in the network depends on the node's ability to take an *intentional stance* – reasons for action that make one's behavior predictable based on their habits or moral values (Dennett, 1989; Dowding, 2008).

This research will not concentrate on the origin of specific agendas or influences of individual companies, as they are not the only ones involved in the co-production process. Rather, it will acknowledge the diversity of goals and interests among non-state actors and the complexity of the technological and social processes that affect AI governance. This approach assumes that any single company or technological innovation does not determine the moral landscape of AI governance. Instead, it focuses on the implications of this co-production for AI

governance while reflecting the likely simultaneous effects of technological advancement on democratic societies and their governing elites.

Outline

In the second chapter, I analyze the moral landscape of AI governance in two regions – the United States and the European Union. First, I reviewed the secondary literature to understand trends in AI. Then, I quantitatively analyzed three corpora of texts in the following way: first, I operationalized the moral landscape and developed a lexicon for AI governance themes; then, I conducted a rigorous time series analysis grounded in advanced natural language processing techniques. Appendices A and B provide a comprehensive breakdown of the data sets (corpora) examined, containing a range of relevant policy texts and documents from these regions. Finally, I compared the quantitative results to the secondary literature. This chapter aims to inform policymakers, researchers, and practitioners, fostering a more informed and responsible approach to AI development and deployment.

Chapter three presents a comparative analysis of two distinct regulatory regimes, the United States and the European Union. Building upon the quantitative insights from the preceding chapter, I analyze the policy documents, dissecting the reasoning and arguments that underpin the governance of artificial intelligence. The research question for this chapter is "How is the moral landscape in AI governance dominated by the same set of non-state actors reflected in the policy-making in two different regulatory regimes?" This chapter will study the reasoning and arguments presented along with and within the policies governing

artificial intelligence by juxtaposing the approaches taken by the US and the EU, focusing predominantly on two distinct themes: AI ethics and existential risk from AI. Through this comparative lens, I seek to inform policymakers, practitioners, and scholars, fostering a more nuanced understanding of the multifaceted challenges inherent in shaping the future of AI.

Chapter four reflects on the evolution of AI ethics and presents a compelling case study that reflects on current approaches to developing AI systems ethically. To answer this challenge, we created an AI Ethics Tool. As AI systems become increasingly integrated into various domains, from healthcare to finance, ethical considerations are raised, but improvements are scarce. The phenomenon of unprecedented self-governance in AI presents an opportunity for impactful intervention. We propose building bridges between human intentionality and AI autonomy by empowering developers and practitioners to assess ethical risks and design ethical algorithms by emphasizing fairness, transparency, and accountability (Floridi et al., 2018; Ryan et al., 2021). We do this by translating ethical principles into actionable design choices, presenting a proof of concept that such an approach can be a powerful venue for stakeholder participation.

In the fifth chapter, I reflect on the moral landscape of AI policy-making in the EU and the US regarding a technology that has recently taken the world by storm: large language models (also referred to as foundation models in the context of EU legislation). I conclude with a discussion and highlight future research directions for AI governance.

Chapter 2

Navigating the Moral Landscape of AI Governance

"After Descartes based his own philosophy upon the discoveries of Galileo, philosophy has seemed condemned to be always one step behind the scientists and their ever more amazing discoveries, whose principles it has strived arduously to discover ex post facto and to fit into some over-all interpretation of the nature of human knowledge."

- Hannah Arendt, The Human Condition

There are compelling arguments that international corporations are increasingly asserting dominance over AI governance, leveraging their vast resources to shape policy outcomes (Roberts et.al., 2024; Cihon, Schuett, & Baum, 2021). This chapter explores the hypothesis that non-state actors, specifically international corporations based in the United States, play a significant role in the moral landscape of Artificial Intelligence (AI), which in turn, profoundly impacts AI policymaking.

To test the hypothesis, I developed a mixed-methods research approach. Firstly, I collected three extensive datasets, each representing the policies of a distinct group of actors: six international corporations based in the United States, the United States federal government, and the European Union. These entities were chosen due to their significant role and influence in the global AI landscape.[8]

[8] According to the OECD, the United States has published 82 AI Policy initiatives, and the European Union has published 63 initiatives. Comparatively, China and Russia have published 22 and 11 initiatives related to AI respectively.

The Timeline

The examined corpora include AI policy initiative texts from 2010 until 2023. This period of 13 years spawned multiple pivotal moments in AI development that shifted the discourse around AI governance. From the early 2010s, when US President Obama's reelection campaign hired a team of machine learning experts to work on big data analytics (O'Neil, 2016), to the 2022 introduction of the global phenomenon ChatGPT, AI governance as a field had to adapt and reflect the dynamically changing nature of its focal point. I separated the data sets into three periods that allowed me to contextualize the individual themes' emergence with some of the defining moments of AI history.

Three separate periods were defined: the early 2010s (2010 - 2014), mid-to-late 2010s (2015 - 2018), and early 2020s (2019 - 2023). These periods of relatively same length allowed me to reflect on the changing moral landscape in the context of technological developments and socio-economic and political responses to AI over time.

The First AI Hype of the Early 2010s: Machine Learning

From 2010 until 2014, the field of artificial intelligence experienced pivotal years. During this period, technological advancements reshaped AI research and applications. From unsupervised learning to the rise of deep neural networks and breakthroughs in reinforcement learning, the early 2010s laid the groundwork for AI and established a momentum that attracted talent and capital.

In 2011, Jeff Dean and Andrew Ng demonstrated that training a face detector is possible without explicitly labeling images containing faces. Their approach involved using large-scale unsupervised learning. Instead of relying on labeled data, which requires a tedious manual process and extensive resources, they "fed" vast amounts of unlabeled images to a neural network. Through this process, the neural network started recognizing human faces by extracting high-level features from the data (Le et. al., 2011; Alom et al., 2018). This highly influential research opened up new possibilities for training neural networks that did not require extensive manual annotation.

The magnitude of this research lies in the demonstration that unsupervised learning could be a powerful tool for feature extraction. AI research moved away from the traditional reliance on handcrafted features by allowing neural networks to learn from unlabeled data. This had profound implications for representation learning, as it suggested that neural networks can autonomously discover meaningful features without explicit supervision (without manual labeling by a researcher). This finding paved the way for subsequent research in unsupervised and self-supervised learning. It highlighted the importance of large-scale data and the potential for neural networks to uncover latent structures in complex data domains (Alom et al., 2018; Samek et al., 2021).

The breakthrough in unsupervised learning for face detection had several practical implications. It allowed for more efficient and scalable face recognition systems. For instance, companies like Facebook and Google adopted similar techniques to improve their photo tagging features. Additionally, this approach

extended beyond faces—unsupervised learning became crucial for other tasks like clustering, anomaly detection, and recommendation systems. Beyond face detection, unsupervised learning techniques found applications in natural language processing (NLP), where word embeddings and topic modeling benefited from large-scale unlabeled text data.

In 2012, Geoffrey Hinton, Alex Krizhevsky, and Ilya Sutskever introduced AlexNet—a deep-learning model that revolutionized AI with a discovery in computer vision. By leveraging convolutional neural networks and deep architectures, AlexNet significantly reduced error rates compared to previous methods (Krizhevsky et al., 2012). Its success further solidified neural networks and deep learning as the most promising subfield of artificial intelligence. The research was significant for their use of graphics processing units (GPUs) instead of state-of-the-art central processing units (CPUs) (Krizhevsky et al., 2012). The impact of AlexNet extended beyond research—it attracted substantial interest, funding, and accelerated progress in computer vision applications.

AlexNet's success underscored the significance of deep architectures. Before AlexNet, shallow models struggled with complex tasks like image classification. The theoretical implication was that depth matters—deeper networks can learn hierarchical representations, capturing intricate patterns and abstractions (Krizhevsky et al., 2012). Researchers realized that deeper networks could more effectively approximate complex functions. This led to innovations like residual networks and transformer architectures, which dominate various AI applications today.

AlexNet's success led to a surge in deep learning adoption across various domains. Computer vision applications, such as image classification, object detection, and segmentation, have significantly improved (Russakovsky, 2013). Companies integrated deep learning models into their products, enhancing features like content recommendation, personalized ads, or medical image analysis. Deep learning models inspired by AlexNet became the backbone for speech recognition, machine translation, and even self-driving cars. The availability of pre-trained deep networks (transfer learning) accelerated practical deployment.

In 2013, a research team at DeepMind, a company known for its cutting-edge AI research, demonstrated that a deep reinforcement learning model could learn the rules of multiple arcade games without any external input (Mnih et al., 2013). The model surpassed human performance in several games, showcasing the power of combining deep neural networks with reinforcement learning. This achievement marked a significant milestone in AI, emphasizing the potential of AI systems to learn complex tasks autonomously. DeepMind's work paved the way for subsequent advancements in reinforcement learning and its applications in various domains.

Deep reinforcement learning (DRL) demonstrated that neural networks could learn optimal policies from raw sensory input. The combination of deep neural networks and reinforcement learning principles challenged the boundaries of what AI systems could achieve. The idea that agents can learn from trial and error without explicit supervision reshaped our understanding of AI capabilities.

DRL transformed gaming and robotics. In gaming, DRL agents achieved superhuman performance in complex games like Go or chess. In robotics, DRL enabled autonomous control, from robotic arms to drones. Companies started using DRL for optimizing supply chains, energy management, and recommendation systems. Beyond gaming, DRL has found applications in finance, i.e., portfolio optimization (Hu et al., 2019); healthcare, i.e., personalized treatment plans (Johnson et al., 2021); and industrial automation, i.e., process optimization (Yang et al., 2022).

The Sobering Reality of Big Data in the Mid-to-late 2010s

The latter half of the 2010s witnessed not only more significant technological leaps but also a significant public commitment to responsible AI in the private sector. Driven by visionary investments, the growing attention of mainstream media, and the emergence of Large Language Models (LLMs), this period furthered AI's transformative impact on society.

In December 2015, a consortium of technology personalities like Elon Musk and Peter Thiel founded OpenAI. Their $1 billion commitment aimed to promote artificial general intelligence research while emphasizing safety and humanity's benefit. OpenAI's establishment promoted the ethical imperative of AGI development. The commitment to openness and safety positioned OpenAI in contrast to existing tech companies, which were criticized for their lack of transparency and responsibility.

In 2015, we also observed the rise and fall of Cambridge Analytica, a political data analytics company behind the Ted Cruz and Donald Trump

campaigns. The Guardian's reporting of the scandal shocked audiences worldwide. The company was developing psychological profiles of Facebook users and their friends to target political advertising better (O'Neil, 2016). The outrage was far beyond the 40 million affected voters in the United States. Facebook's CEO, Mark Zuckerberg, later testified in front of the US Congress, and the company had to pay a $5 billion fine for its role in the data harvesting scheme.

In March 2016, AlphaGo—a neural network model developed by DeepMind—stunned the world by defeating Lee Sedol, the reigning Go champion, in a historic five-game match. AlphaGo showcased the power of deep learning and reinforcement learning, moving beyond mere computational "brute force" approaches. AI could intuit, strategize, and outperform human experts (Silver et al., 2016). AlphaGo's success validated neural networks and reinforced learning. Bidirectional training, a key component, allowed AlphaGo to understand context more effectively. Beyond Go, AlphaGo's victory further solidified global interest in AI's potential applications.

At the close of 2018, Google's BERT (Bidirectional Encoder Representations from Transformers) revolutionized natural language processing. BERT's bidirectional training, capturing both preceding and following context, addressed the limitations of earlier models. This breakthrough would soon cause a seismic shift with the advent of Large Language Models (LLMs) (Devlin et al., 2018).

BERT's bidirectional approach transformed NLP, enabling nuanced understanding. LLMs shifted the paradigm from task-specific models to pre-trained, adaptable architectures. BERT's impact extended beyond NLP, influencing diverse fields like information retrieval and sentiment analysis. LLMs became foundational for chatbots such as ChatGPT, virtual assistants, and content generation.

Google's two controversial projects, Dragonfly and Maven, fueled critics in 2018 who called for more transparency and accountability from the big tech but simultaneously represented a positive example of employee activism. Both the development of a censored search engine for China (in the case of the project Dragonfly) and the Pentagon program implementing artificial intelligence to interpret video imagery and potentially improve the targeting of drone strikes (project Maven) were reportedly shut down after internal protests (Shane & Wakabayashi, 2018; Gallagher, 2018; Amnesty International, 2020; Copestake, 2018; Pellerin, 2017).

These defining moments of the second half of the 2010s gave rise to a much more complex power dynamic. They challenged the technologically deterministic language that was so prominent in the decade's early years. Technology became increasingly morally and politically charged; technological companies could no longer claim to be outside the social and political realm. AI evolved from an obscure set of equations to a socio-technical phenomenon.

The second AI Hype in the Early 2020s: Large Language Models

From 2019 to 2023, we have witnessed exponential growth in artificial intelligence. This period was marked by breakthroughs, controversies, and ethical dilemmas that echoed across academia, industry, and society.

GPT-2 (Generative Pretrained Transformer 2) attracted attention in natural language understanding research and catalyzed the rise of large-scale language models. GPT-2 demonstrated the power of unsupervised learning on massive text corpora, revealing the potential of neural networks to capture intricate linguistic patterns. Built on the Transformer architecture, GPT-2 harnessed attention mechanisms to capture long-range dependencies, enabling context-aware predictions (Radford et al., 2019).

GPT-2 was part of the broader trend toward pretraining large neural networks on diverse data sources, followed by fine-tuning for specific tasks. Researchers grappled with the balance between model size, computational resources, and performance gains. GPT-2 excelled in text generation, translation, summarization, and question answering.

The *Stochastic Parrots* paper, authored by Timnit Gebru and colleagues (2020), highlighted ethical challenges in Bias and Fairness. The paper emphasized biases in training data and their impact on downstream applications, raising awareness about the environmental footprint of large language models.

The paper and its reception underscored the tension between corporate interests and responsible AI research. Corporate control over research became a

focal point of discussion. Google's refusal to publish the paper led to concerns about transparency and accountability. The consequent ousting of Dr. Gebru and Margaret Mitchell highlighted the need for an inclusive and transparent AI community.

In 2022, conversational breakthroughs of ChatGPT, built on GPT-3.5, pushed the boundaries of conversational AI (OpenAI, 2022). ChatGPT showed substantial progress in understanding natural language and engaging in interactive conversations. It bridged the gap between human and machine communication, leveraging advancements made by GPT-2 and GPT-3. ChatGPT's interactive conversations on almost any subject brought the company unprecedented success: the chatbot is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history (Hu, 2023). Researchers explored ways to improve model accuracy and mitigate biases. However, concerns arose about its potential misuse, including spreading misinformation and producing harmful content (Bogost, 2022; Roose, 2023).

Data and Methods

The data collection involved gathering policy documents from multiple sources to create three distinct corpora.[9] The period for all three corpora was set from 2010 to the present, ensuring more than a decade-long perspective on the evolution of AI governance. For the European Union and United States data, the primary source was the OECD AI Policy Database (OECD, 2023). This

[9] Appendices A, B, and C provide the full list of policy initiatives.

comprehensive database provides information on AI policies and regulations. However, to ensure a complete representation of the policy landscape, we supplemented this data with additional important policies and laws that were not included in the OECD database.

The types of policy initiatives present in the EU corpus include research and development initiatives like The Human Brain Project of 2013 (European Commission, 2013), which promoted the use of most advanced AI techniques for accelerating progress in neurosciences, or the European Parliament's Resolution on Digitising European Industry of 2017, which highlights the need to "establish leadership in digital industrial value chains and key technologies such as 5G, quantum technologies, high-performance computing, artificial intelligence, cloud computing, big data analytics, the Internet of Things (IoT), robotics, automation (including Highly-Automated Driving) and Distributed Ledger Technology" (European Parliament, 2016). The EU corpus also includes essential regulations such as the General Data Protection Regulation of 2016, which aimed to empower EU citizens and EU-based individuals to take control over their data and simplify international businesses' regulatory environment by unifying the regulation within the EU market. Lastly, the corpus includes regulatory frameworks such as the Digital Markets Act of 2022, which addresses unfair and uncontestable digital markets where tech companies in entrenched and durable positions act as gatekeepers, preventing new businesses from emerging and competing (Yasar et al., 2024).

Similarly, the US corpus includes initiatives such as the National Robotics Initiative of 2011, a pioneering effort that has significantly advanced the science of robot integration globally (National Science Foundation, 2011). The Big Data to Knowledge national R&D program of 2016, supported innovation and transformative approaches to maximize the adoption and accelerate the utility of big data and data science in biomedical research (National Institute of Health, 2016). The corpus also includes investment and strategic announcements by the Defense Advanced Research Projects Agency (DARPA), such as the Explainable AI Program in 2018, aiming to create a suite of ML techniques that produce more explainable models (that enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (Kejriwal, 2021) while maintaining a high learning performance and, or the AI Next Campaign in 2018, a multi-year investment of more than $2 billion on AI R&D in a portfolio of some 50 new and existing programs. Finally, influential frameworks developed by federal government agencies to address systemic risks and propose voluntary practices and norms, such as the National Institute of Standards and Technology (NIST) AI Risk Management Framework of 2023 or the Blueprint for AI Bill of Rights from late 2022, are also featured.

To collect similar data representing the non-state actors, I conducted an extensive internet search and screened the companies' websites to gather relevant policy documents. To ensure a representative dataset, I also utilized the Internet Archive service "Wayback Machine." This tool allowed me to access historical

versions of these websites, enabling me to collect data that spanned the same period as the data from the US and EU regulatory bodies.

The policy initiatives in the non-state corpora include Amazon's Framework to mitigate bias and improve outcomes in the new age of AI (2023) or the company's partnership with the NSF Program on Fairness in Artificial Intelligence in Collaboration (2019). Microsoft's self-governing efforts include a position paper on Social and Ethical Implications of Autonomous Experimentation in AI (Bird et al., 2016), the AI Whitepaper for the Age of Intelligence (2018), or The Future Computed: Artificial Intelligence and Its Role in Society (2018) book. OpenAI is represented in the corpus by its mission statement (2015) and a policy paper on Concrete AI Safety Problems (2016). For instance, Google's corporate policy initiatives are the Recommendations for Regulating AI (2018). Apple, the notoriously least engaged and least publicized of the studied tech companies, limits its public stance on AI governance to its Ethics and Compliance statement, overseen by the Board of Directors. Their minimalistic communication has, however, expanded through their membership in the Partnership on AI (Shead, 2017).

I included relevant policy documents from the Partnership on AI, which is an organization established in 2016 by Amazon, Facebook, Google, DeepMind, Microsoft, and IBM, with interim co-chairs Eric Horvitz of Microsoft Research and Mustafa Suleyman of DeepMind (Hern, 2016; Waters, 2016; Bindi, 2016; Rubin, 2016), and joined by Apple soon after in 2017. As of April 2024, the board

of directors had representatives from each of the studied private companies (Amazon, Apple, Google (DeepMind), Microsoft, Meta, and OpenAI.)

Developing the AI Governance Lexicon

I employed a multi-step process to develop a lexicon of themes that have emerged in the discourse surrounding AI and AI governance. The final lexicon encompasses themes not only from policy debates but also from popular culture, *grey literature*, various scientific disciplines, and journalistic coverage of AI.

Developing the lexicon involved a systematic approach, starting with creating word lists associated with specific themes. The AI Moral Landscape Lexicon (see the complete lexicon in Appendix C) emerged from a thorough process that combined theoretical frameworks, empirical research, and ongoing refinement. I aimed for the lexicon to serve as a valuable tool for navigating the complex terrain of AI governance.

I operationalized Abend's moral background as a framework for theme identification (see Table 1). This approach ensured that the lexicon captured a comprehensive set of words for each theme, reflecting the different layers of the moral landscape. This framework provided a structured way to explore different moral dimensions. Through an inductive process, I identified key themes related to AI governance. These themes encompassed various aspects, from ethical considerations to legal and policy implications.

For each theme, I compiled a list of relevant words and terms. These words represented the vocabulary associated with different facets of AI. The goal

was to capture a comprehensive set of terms that spanned the moral landscape, ensuring that the lexicon covered diverse perspectives and nuances.

Table 1. Operationalized Moral Landscape for AI Governance

Landscape Dimension	**Conceptualized Dimension**	**Operationalized Dimension**
Grounding	What are the grounding principles/normative theories?	What are the moral theories related to AI in English language research papers, news articles, and popular culture, as evidenced by the frequency and context of related keywords?
Method and Argument	What evidence/ nature of arguments is used? What is the evidence used for moral reasoning around AI?	What are the types of evidence and nature of arguments used for moral reasoning around AI in English language research papers, news articles, and popular culture, as evidenced by the frequency and context of related keywords?

Object of Evaluation	What are the aspects of AI that we subject to moral evaluation?	What are the aspects of AI that are subject to moral evaluation in English language research papers, news articles, and popular culture, as evidenced by the frequency and context of related keywords?
Repertoire of Concepts	What are the moral concepts that we see in AI?	What are the concepts that allow us to morally evaluate AI in English language research papers, news articles, and popular culture, as evidenced by the frequency and context of related keywords?
Metaethics	What is the nature of moral statements around AI? Do moral agents believe in their objectivity?	Not suitable for quantitative analysis.
Metaphysics	What are the social facts and social groups of AI? What are the relevant social entities and phenomena?	Not suitable for quantitative analysis.

I then engaged in an iterative refinement process. Continually, I expanded upon these themes by combining my domain knowledge and various sources,

which developed the lexicon and ensured its thoroughness and relevance to the current AI governance landscape. I consulted various sources, including academic literature, policy documents, and expert opinions. This allowed me to enrich the lexicon by adding contextually relevant terms. I aimed to create a resource that reflected the evolving landscape of AI governance. As new developments occurred, I updated the lexicon accordingly. Rigorously reviewing and refining the lexicon, my ambition is for it to be a valuable reference for researchers, policymakers, and practitioners (see Table 2 for an illustration of the Moral Landscape Lexicon).

Theme	Related Keywords
Data_Governance	anonymization, consent, data, data access controls, data audits, data breach, data ethics, data ethics boards, data governance frameworks, data minimization, data ownership, data portability, data privacy impact assessments, data provenance, data protection, data quality, data retention, data sharing, data sovereignty, data stewardship, data localization, privacy
Defense_and_State_Security	adversarial attacks, ai in defense, arms control, autonomous drones, counterterrorism, critical infrastructure protection, cybersecurity, cyberwarfare, defense, defense policy, dual-use technology, geopolitical stability, intelligence sharing, military applications, military readiness, national resilience, national security, nuclear deterrence, secure ai deployment, secure development, strategic alliance, surveillance

Table 2. Related keywords (concept search terms) in the AI Moral Landscape Lexicon (see complete lexicon in appendix C).

AI Policy Language Assessment Utilizing the AI Moral Landscape Lexicon

I analyzed the three policy corpora using various computational approaches. The primary objective of this analysis was to determine whether the themes predominantly present in the non-state actor corpus align with those in the other two corpora (Schuelke-Leech, & Jordan, & Barry, 2019). This alignment, if present, would provide empirical support for the hypothesis that non-state actors may be the dominant force shaping the moral landscape of AI. Once the lexicon was established, I prepared it for data analysis using standard Natural Language Processing (NLP) techniques.

This preparation involved several steps. Firstly, I removed so-called “stopwords,” which consist of eliminating common words (e.g., “and,” “the,” and “is”), as well as unnecessary characters, spaces, and punctuation from the text that do not contribute much information for the analysis. Then I tokenized (which involves breaking down the text into individual words or tokens) and stemmed the text (which reduces words to their root form; for example, "running," "runs," and "ran" would all be reduced to the stem "run*"). After the data preparation, I proceeded to analyze the three corpora using the prepared lexicon.

Figure 1. Wordcloud illustration of the word frequency analysis of the Non-state actor's AI Policy Initiatives from the early 2010s until 2023.

Figure 2. Wordcloud illustration of the word frequency analysis of the AI Policy Initiatives in the European Union from the early 2010s until 2023.

In the language analysis phase of this research, I first conducted a simple word frequency analysis generating word clouds. A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance in the corpus. This technique provided a quick and intuitive understanding of the most prominent words and themes in each dataset (see figures 1-3).

Following the word frequency analysis, I utilized the lexicon to search for the presence of specific themes within the corpora. This involved scanning the text data for the words associated with each theme and quantifying their occurrence. This approach allowed us to identify and compare the prevalence of various themes across the three datasets.

Figure 3. Wordcloud illustration of the word frequency analysis of the AI Policy Initiatives in the United States from the early 2010s until 2023.

I leveraged several programming libraries for the analysis. The Natural Language Toolkit (NLTK) is a leading platform for building Python programs with natural language data (Hardeniya et al., 2016). It provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet, along with a suite of text-processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Pandas is a software library for Python that provides data manipulation and analysis capabilities. It offers data structures and operations for manipulating numerical tables and time series, making it an ideal tool for handling and analyzing large datasets. Matplotlib is a plotting library for Python. It provides an object-oriented API for embedding plots into applications. In this research project, I used Matplotlib to create visualizations, which helped illustrate the findings of the analysis.

After the initial assessment of the full corpora, I studied the prevalence of individual themes via search terms (keywords). I aimed to identify dominant themes within each text corpus in raw numbers and then normalized the results to percentage values to allow for a comparison across unevenly sized corpora. This was crucial for two reasons. First, policy documents grappling with AI were much scarcer in the early 2010s than by the end of the same decade. Second, there were also significant disparities in the amounts and lengths of documents across the three actors, particularly the non-state actors, whose publications were much more often gray literature aimed at popular and journalistic consumption than verbose EU-style legislation.

Results & Discussion

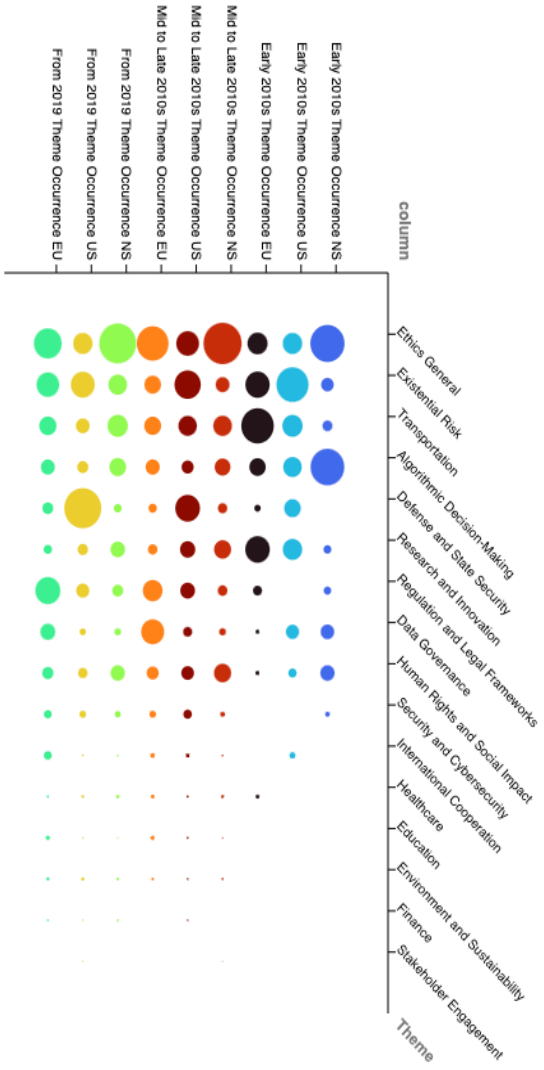

Figure 4. Theme occurrence in the three analyzed corpora is separated into three timeframes: early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s

(containing the period 2019 - 2023).

As seen in Figure 4, the most prominent themes overall were Ethics General and Existential Risk, with the former steadily the most dominant theme of the moral landscape in non-state actors' AI policy initiatives and the latter a significant theme across the EU and even more significantly the US corpus.

Defense and State Security were, by a large margin, the most prominent themes in the United States AI Policy initiatives in the mid-to-late 2010s and even more so in the early 2020s (see Figure 6 for a timeline). Transportation is a significant theme, particularly for EU policymakers, perhaps in response to ongoing developments in the autonomous vehicle industry. Algorithmic decision-making, a theme encompassing technical debates such as model performance, evaluation, and monitoring strategies, was the prime focus of non-state actors in the early 2010s era of the first machine learning hype but remained an important theme across the corpora in all three periods.

Important insights and context will only be recovered by examining the corpora and setting of the moral landscape in the context of technological and socio-political realities. In the following sections, I dive deeper into the narrative of the technology actors, both the companies and the artifacts, to gain a deeper understanding of the potential motivations for the particular makeup of the moral landscapes of each period. I will present a qualitative analysis of the AI Ethics and Existential Risk themes to untangle the particulars and reflect on the hypothesis for their dominance within the landscape.

Figure 5. Theme occurrence in AI policy initiatives in the EU, the US, and the non-state actors in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s (containing the period 2019 - 2023).

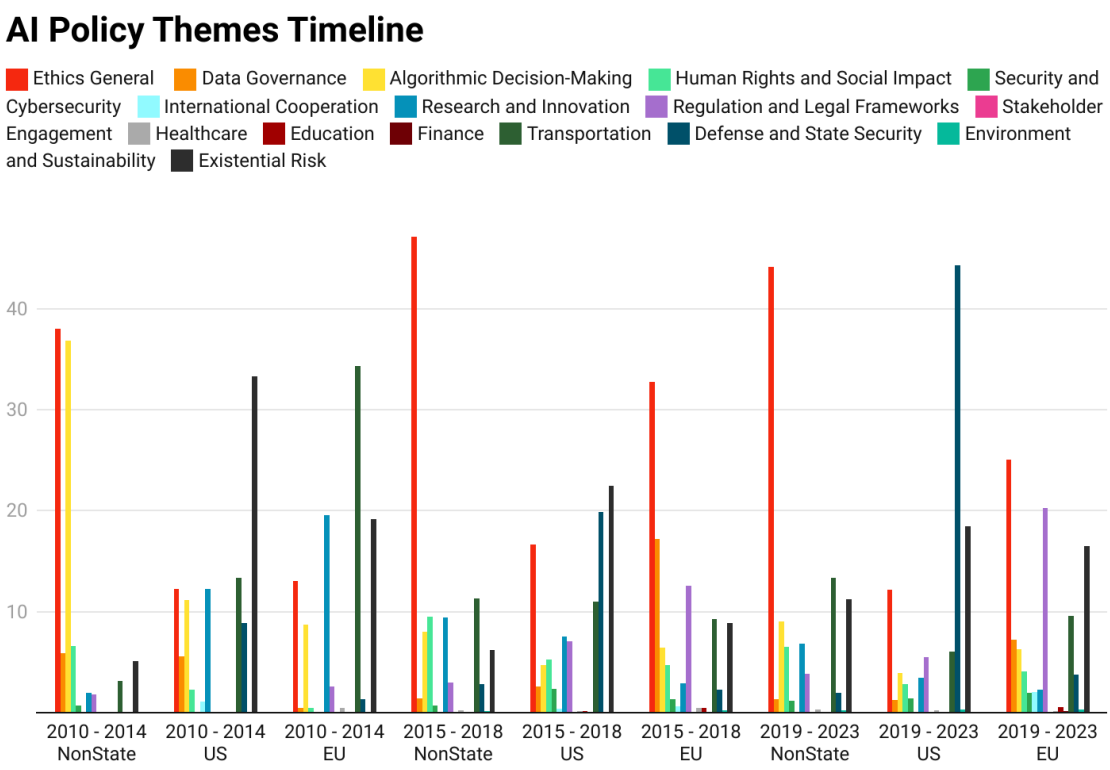

Figure 6. A timeline of the theme occurrence in AI policy initiatives in the EU, the US, and the Non-state actors in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s (containing the period 2019 - 2023).

The Power-Knowledge and the AI ExRisk

The "Existential Risk" (Existential Risk theme or Ex-Risk theme) represents debates over the catastrophic risks associated with AI, particularly *strong AI* or *AGI* (Bostrom, 2016; Cave & ÓhÉigeartaigh, 2018). Existential risks are threats that could endanger the existence of humanity or cause extreme and irreversible adverse outcomes (Namdar, & Pölzler, & Ord, 2020). One critical concern is the development of *autonomous weapons systems*. These AI-powered weapons, capable of making life-or-death decisions without human intervention,

pose a grave risk to global security and stability (Bode, 2023). Additionally, the *race dynamics* surrounding AI—where nations, organizations, and researchers compete for breakthroughs—can inadvertently escalate risks (Kissinger & Allison, 2023; Hirsh, 2023). Furthermore, the *control problem* challenges ensuring that *superintelligent AI* systems remain aligned with human values and goals. Bostrom's *orthogonality* thesis states that intelligence and values are not inherently linked, necessitating deliberate efforts to align AI with our ethical principles (Bostrom, 2012b), followed by *Maxipok principle*—maximizing the probability of humanity's survival and flourishing.[10]

The threat of *unaligned artificial intelligence*—systems whose goals diverge from human values—requires rigorous research in *value alignment*, the process of ensuring AI, particularly more sophisticated, general AI, shares our ethical principles (Russell & Norvig, 2021; Russell, 2020). Moreover, this theme also discusses the impact of AI on employment. *AI-induced unemployment* could destabilize societies and, if not addressed, exacerbate existing inequalities and ultimately threaten democracy (Gordon & Gunkel, 2024).

Perhaps surprisingly low is the focus on an existential threat in the early 2010s non-state actors' policies, who are often suspected of promoting this type of *agenda* to "throw off" policymakers from their tracks and instead focus on hypothetical scenarios rather than present issues (Schermer, 2017; Goldman 2023; Jindal, 2023). In comparison, in the early 2010s, Ex-Risk was the most prominent theme in US AI policies and one of the two most prominent still in the mid-to-late

[10] "Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe." (Bostrom, 2012a).

2010s. In the European policy initiatives, the Ex-Risk theme occupies a significant portion of all policy efforts as well; nineteen percent of all policy efforts in the early 2010s contain the theme of Ex-Risk, nine percent in the mid-to-late 2010s, and almost seventeen percent in the early 2020s.

The slow uptake of Ex-Risk by non-state actors may have two explanations; in the early 2010s, the already negative portrayal of AI in popular culture may have been seen as a potential obstacle to the deployment of AI. From the position of established, dominant experts, it made more sense to remain indifferent to ExRisk and further establish a dominant position by focusing on technical *finesse* (Algorithmic Decision-Making is a particularly prominent theme in the early 2010s moral landscape for non-state actors) over hypothetical scenarios that could lead to an abrupt government intervention.

In line with the first explanation is a second possible answer—the early 2010s concern in the US government was framed by non-state actors as evidence of lacking expertise. Brushing off the concern of the US Senator over the fear of *Terminator* might have been an effective strategy to solidify epistemic authority and fuel the narrative of technocracy as the superior way of governing technologically advanced societies.

Did the language of Ex-Risk become a tool for tech CEOs to paralyze democratic deliberations over AI governance and insert themselves into positions of power? Or did the thousand-year history of fascination with AI fuel the tech companies' legitimacy to assert their agenda in the political domain? Perhaps both are true at the same time. King and Hayes proposed that experts in hazardous

industries limit how regulators regulate risks and constrain the potential of regulatory outcomes through their unique epistemic position of power (King & Hayes, 2017). This new form of regulatory capture, they argue, stems from Foucauldian power-knowledge. They study the performance of power-knowledge in relation to risk regulation by and around regulators. They draw the understanding of the power-knowledge relationship from Foucault's notion of power: "Power is relational and operates through elements of apparatus, including discourses, strategies, technologies, institutions, regulatory decisions, laws, administrative measures, scientific statements, philosophical stances, etc. The apparatus is embedded in the exercise of power, linked with and supported by types of knowledge (Foucault 1980, 196)" (King & Hayes, 2017).

Tech companies' growing interest in the theme of Ex-Risk, could be hence understood as a rational economic actor's response to an emerging opportunity, or, through the actor-network and co-production lens, as a result of multiple concurring effects of the network. While one set of actors gained legitimacy, others were disenfranchised by the cultural and socio-economic narratives of what it means to govern AI. The power-knowledge performance of non-state actors from the early 2010s captured more than just regulators; the promise of abundance and the threat of an ultimate end became a carrot and stick to Western societies on both shores of the Atlantic Ocean.

Moreover, the discoveries from the early 2010s ignited human imagination beyond strictly technical circles. A machine's ability to "recognize" patterns or even human faces learning by itself was much more sensational to mainstream

media consumers than the software engineering world. Indeed, even the revolutionary AlexNet had a predecessor, LeNet, in 1995 (LeCun et al., 1995) that technical audiences could recall. Yet, the imagery of a machine that can learn about the world without explicit human assistance confirmed the popular culture narratives stemming from thousands of years of human fascination with artificial agents (Cave et al., 2018).

Hence, unsurprisingly, the inclination to address existential threats and adopt risk-based approaches to AI governance mostly stems from elected regulators, not Silicon Valley's scheming CEOs. However, this powerful narrative still contributes to the lack of democratic oversight in AI governance. The academic community must keep challenging the anthropomorphization of AI and reject technological determinism. The power-knowledge performance of tech companies is perhaps best illustrated by the following incongruity: AI is a technology that is only possible due to the global, collective creation of enormous datasets, yet its development is hidden behind walls of elaborate research campuses in California, shrouded in a haze of numbers seemingly without meaning.

The Desire for AI Ethics

The theme of AI Ethics has steadily been Silicon Valley's main concern since the early 2010s. Conversely, it is an afterthought for many of the policy initiatives in the United States. In the European Union, AI Ethics became the most important topic in the mid-to-late 2010s and remained the focal point in the early 2020s (see Figure 6 for the timeline comparison).

Some researchers put AI Ethics in contrast to the Ex-Risk theme (Cave & ÓhÉigeartaigh, 2019). While one predominantly challenges existing practices and grapples with ethical issues in present-day software, the other is sometimes criticized for not paying enough attention to contemporary issues like social justice and being too preoccupied with hypothetical scenarios such as superintelligence (Prunkl & Whittlestone, 2020). However, there are important overlaps between the two themes. Moral values and their translation or alignment with software development practices and the technology product are crucial problems in both themes.

The Ethics General theme encompasses the prominent debates over AI ethics. This facet of the AI moral landscape contains a range of principles and practices to ensure responsible development, deployment, and use of AI systems (Vakkuri, 2021; Morley, 2021; Martinho, 2022). Key aspects include *accountability, transparency, fairness, and minimizing harm*.[11] An *AI ethics committee* can play a pivotal role in evaluating and guiding ethical decisions (Schuett et al., 2024), while *impact assessments* (UNESCO, 2023) help assess potential effects on society. *Algorithmic bias, privacy by design,* and *stakeholder engagement* are essential components in building ethical AI systems (Gebru, 2020). *Civil society engagement* and other considerations regarding threats to democracy are also notable (Shattuck et al., 2022). Some aspects of the AI Ethics theme also overlap with the Existential Risk category, namely where aligning

[11] From the point of view of western normative ethics, deontological and consequentialist, as well as virtue ethics theories are represented in the mainstream AI Ethics debates, though the first two have significantly stronger representation (Hagendorff, 2020; Hagendorff, 2022).

values with technology or *dual-use technologies* are discussed (Urbina et al., 2022).

Why is AI Ethics the focal point for non-state actors? In most cases, private companies' economic motivation will be to minimize state intervention. They will attempt to establish strategic partnerships with the public sector to regain control and prevent state regulation through sufficient self-governance. By signaling good practices, companies not only appease policymakers but also establish trust with the general public (Roski et al., 2021).

The wish for moral AI is also, to some extent, motivated by a deeper Kantian sense of praiseworthiness (McCarty, 2009). Kant argues that to be worthy of praise for one's actions, one's motivations must be following one's moral code.[12] In other words, tech companies are preoccupied with aligning AI with their moral values to ensure that its behavior is correct and that the machine arrives at its choice motivated by the approved set of morals. The urgency to build praiseworthy AI is most apparent in policies where value alignment overlaps with Ex-Risk themes, such as in the "Planning for AGI and Beyond" mission statement by Sam Altman, the CEO of OpenAI (2023), where he argues for a cautious development of superhuman AI aligned with human values as "the most important project in human history" (Altman, 2023).

[12] The moral code, or as Kant calls it, a maxim, is only permissible if the same code could be applied universally, or as Kant puts it, the *Categorical Imperative* is to "Act so that through your maxims you could be a legislator of universal laws" (Kant, 1785).

Chapter 3

Moral Landscape of the American and European AI Governance: A Comparative Study

"The difficulty lies not so much in developing new ideas as in escaping from old ones."

- John Maynard Keynes, The General Theory of Employment, Interest and Money

In the previous chapters, I introduced the concept of the moral landscape of AI governance. I proposed a mixed-method analysis of the moral landscape utilizing natural language analysis. I presented the quantitative results in the context of AI's development in the past two decades. I then analyzed two themes with a strong, steady presence across the corpora over time: existential risk and AI ethics. I propose a rationale for each theme focusing on the non-state actors in AI governance. In this chapter, I focus on the US and the EU's approaches to AI governance and the two regions' response to the self-governance of Silicon Valley.

Beyond Artificial Intelligence: the Regulatory Approaches of the United States and the European Union

The United States and the European Union regulatory approaches differ due to important historical, cultural, and economic differences that extend beyond their approach to technologies.

The United States' regulatory approach is heavily influenced by its emphasis on individualism and market freedom (Summers, 2001). There is a strong preference for minimal government intervention in business, reflecting a broader cultural value of personal responsibility and an entrepreneurial spirit (Summers, 2001; Redmond, 2004; Farrell, 2003). Historically, the U.S. regulatory system has evolved from a foundation of federalism, where states retain significant powers (Rosenthal & Joseph, 2017). This historical context has led to a decentralized regulatory approach, with regulations varying across states (Henrikson, 2016). The U.S. economy is characterized by its dynamic and competitive nature, with a strong focus on innovation and technological advancement (Greif & Nye & Kiesling, 2020). This economic environment supports a flexible and adaptive regulatory approach, allowing businesses to thrive with fewer constraints (Craig et al., 2017; Engler, 2023).

The EU's regulatory framework is influenced by a cultural focus on social welfare and collective responsibility, which is demonstrated through the precautionary principle, which prioritizes public health and environmental protection over economic freedom (Movius & Krup, 2009; Atkinson & Scott,

2009; Borrás, 2003). The EU's economy, while competitive, places a strong emphasis on stability and sustainability (Orbie & Babarinde, 2013). This is evident in the comprehensive regulatory frameworks that aim to create a fair playing field across member states and ensure long-term economic and social stability, providing security and predictability (Prove, 2018; Takács, 2014). The EU's regulatory system has been shaped by its history of integration and cooperation among diverse member states. The necessity to standardize regulations across different legal and political systems has led to a more unified regulatory framework (Matthijs & Parsons, 2022).

An Overview of AI Policy Initiatives in the United States

The United States has several federal agencies that are responsible for regulating the development and use of artificial intelligence, including the Federal Trade Commission (FTC), the Department of Commerce, and the Department of Defense. These agencies enforce relevant existing laws and regulations and provide guidance on the responsible development and use of artificial intelligence. Additionally, the government has established The National AI Advisory Committee (NAIAC) (National Artificial Intelligence Advisory Committee, 2023), and the Office of Science and Technology Policy (OSTP) supports the National Artificial Intelligence Initiative Office (Office of Science and Technology Policy, 2022). The National Institute of Standards and Technology (NIST), the Department of Commerce agency, has released several documents to inform the position of the United States on AI development and to engage the relevant parties (notably Tabassi, 2023). Along with these specialized

groups, the American Congress has introduced several bills, particularly the American Senate, which has held several congressional hearings with prominent executives of tech companies.

Several existing laws and regulations regulate the development and use of artificial intelligence in the United States. One key law that affects this is the Federal Trade Commission Act, which prohibits deceptive or unfair practices in the marketplace. Additionally, the United States has some sector-specific laws and regulations that address the use of AI in specific industries, such as the Health Insurance Portability and Accountability Act (HIPAA).

The US approach to AI governance has allowed tech companies to set their ethical standards and practices. However, several initiatives address specific issues related to AI. In October 2016, President Obama's White House issued a document titled Preparing for the Future of Artificial Intelligence (The White House, 2016a), which provided an overview of the current state, applications, and implications of AI, as well as recommendations for future actions and investments by the federal government and other stakeholders. The document was accompanied by a supplementary R&D document, the National Artificial Intelligence Research and Development Strategic Plan (The White House, 2016b), which outlined the strategic vision, priorities, and coordination mechanisms for federal AI R&D. The plan identified seven strategies for advancing the scientific and technological foundations of AI, ensuring the safety and security of AI systems, developing shared public datasets and environments for AI training and testing, measuring, and evaluating AI systems through

standards and benchmarks, understanding the national AI R&D workforce needs, and expanding public-private partnerships to accelerate advances in AI. The plan also emphasized the importance of addressing the ethical, legal, and societal implications of AI and fostering public trust and engagement in AI.

Executive Order 13859, issued by President Trump on February 11, 2019 (The White House, 2019), establishes the American AI Initiative as a coordinated Federal Government strategy to sustain and enhance the US leadership position in AI research and development (R&D) and deployment. The Initiative is informed by five principles that aim to foster technological innovation, reduce regulatory barriers, develop workforce competencies, safeguard national security interests, and uphold public trust and ethical values in AI. The Initiative also delineates the roles and responsibilities of various agencies to implement and oversee the Initiative while promoting collaboration with industry, academia, and international partners.

Finally, considered to be the most comprehensive piece of governance by the United States regarding AI, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (The White House, 2023) was issued by President Biden on October 30, 2023. The executive order establishes a comprehensive and coordinated Federal Government strategy to govern and advance AI under eight guiding principles and priorities. The order sets new standards for AI safety and security, such as requiring developers of high-risk AI systems to share their safety test results and other critical information with the U.S. government and developing standards, tools, and tests to ensure that

AI systems are safe, secure, and trustworthy. The order also protects Americans' privacy, civil rights, and civil liberties by directing agencies to assess and mitigate the potential harms of AI systems, such as bias, discrimination, and disinformation, and to enforce existing laws and regulations that protect consumers, workers, and communities from AI-related harms. The order further promotes innovation and competition in AI by supporting long-term research, increasing access to data and computing resources, reducing regulatory barriers, and expanding public-private partnerships. The order also advances American leadership in AI by engaging with international allies and partners, promoting democratic values and human rights, and protecting national security interests and critical infrastructure from AI threats. The order assigns specific roles and responsibilities to various agencies and interagency bodies to implement and oversee the order and requires regular reporting and evaluation of its progress and outcomes.

Aside from the executive orders, the OECD.ai Policy Observatory records 80 sector-specific policies on AI to date, such as the National Institute of Standards and Technology (NIST) AI Risk Management Framework (the NIST AI RMF).[13] The NIST AI RMF was first published in January 2023. According to its authors, it is a "guidance document for voluntary use by organizations designing, developing, deploying or using AI systems to help manage the many risks of AI technologies." (NIST,2023). NIST received a mandate from Congress in 2020 to lead the effort to develop the framework, which was conducted in close

[13] See Appendix A for a full list.

collaboration with the public and private sectors through a series of workshops. The framework builds on NIST's previous risk management frameworks, particularly cybersecurity. The framework is designed to be implemented by developers and practitioners working in organizations across the AI community with a set of practical steps that aim to support the development of more trustworthy AI. While the RMF is intended to assist technical audiences, it does not dismiss the importance of organizational culture, leadership, and governance; it states that "AI systems are inherently socio-technical in nature" and introduces the concept of *AI Actor*, which spawns across the initial stages of conceptualization of AI project to end users of the technology (Tabassi, 2023).

The AI Bill of Rights Blueprint is a sector-specific, non-binding policy effort issued by the White House Office of Science and Technology Policy (OSTP) in October 2022. The AI Bill of Rights centers on individuals' national values and rights, which AI systems should reflect and preserve. The framework applies to all automated systems that have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services. Five principles are defined in this framework: Safe and Effective Systems ("You should be protected from unsafe or ineffective systems."); Algorithmic Discrimination Protections ("You should not face discrimination by algorithms and systems should be used and designed equitably."); Data Privacy ("You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used."); Notice and Explanation("You should know that an automated system is being used and

understand how and why it contributes to outcomes that impact you."); Human Alternatives, Consideration, and Fallback ("You should be able to opt-out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.") (OSTP, 2022).

Other notable efforts in the United States include the Federal Trade Commission, which published a policy statement in November 2022, *Regarding the Scope of Unfair Methods of Competition Under Section 5 of the Federal Trade Commission Act*, arguing additional enforcement powers in antitrust cases beyond the established rules of competition, which implies the effort of the FTC to become more aggressive in enforcing competition rules specifically in the digital markets. These efforts, if successful, could mirror stricter competition rules enforcement in the digital sector of the European Union under the Digital Markets Act (2023). Indeed, the FTC seems to have developed a more interventionist stance in the past decade, when the first series of antitrust lawsuits filed with the Department of Justice claiming that Big Tech companies have engaged in anti-competitive methods and behaviors concerning social media platforms (Facebook, Inc., FTC v., 2021, refilled in 2022), search engine (Google Inc., FTC v., 2013 and Google Inc., FTC v.,2020), app stores (Apple Inc., FTC v., 2014, Epic Games v. Apple, 2020), or online advertising (Google Inc., FTV v., 2023).

An Overview of AI Policy Initiatives in the European Union

The European Union governs artificial intelligence through legislative proposals by the executive branch, the European Commission, the legislative branches, the European Parliament and the Council of the EU, and through a

network of agencies such as the European Data Protection Board (EDPB), The European Data Protection Supervisor (EDPS), The European Union Agency for Cybersecurity (ENISA), the Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE), and the High-Level Expert Group on Artificial Intelligence (HLEG AI).

EU leadership gained momentum in AI-related policy initiatives in the European Union in 2016 with the introduction of the final version of the General Data Protection Regulation (GDPR) in April 2016 (effective May 25, 2018). The GDPR replaced the earlier Data Protection Directive. It regulates how companies process and use personal data collected from consumers online. The GDPR was designed to adapt to our interconnected world, where data serves as a common currency, and it modernized privacy principles to fit contemporary technologies and practices better. It is a crucial component of EU privacy and human rights law, emphasizing individuals' control over their personal information and simplifying regulations for international business.

In February 2020, the European Commission introduced a White Paper on AI together with the report "The Safety and Liability Aspects of AI" and the communication "A European Strategy for Data" as part of a wide package on Artificial Intelligence. The White Paper was accompanied by an open call for feedback and signaled the EC's intention to address the challenges of AI in the EU.

The European Commission (EC) unveiled its ambitious agenda for Europe's digital transformation, which encompassed a new industrial strategy to

foster the competitiveness and resilience of the EU's industry in the global market, a digital finance package to enable innovation and reduce market fragmentation in the financial sector, and new rules for tax transparency on digital platforms to ensure fair taxation in the digital economy. The EC also put forward the Digital Services Act (DSA) and the Digital Markets Act (DMA), two landmark regulations that aim to create a safer and more open digital space for consumers and businesses and to ensure fair and contestable markets for online platforms. Moreover, the EC adopted a new EU Cybersecurity Strategy, which sets out a comprehensive framework to enhance the EU's collective resilience, deterrence, and response to cyber threats and promote a global and open cyberspace.

In 2021, the EC presented Rules for Excellence and Trust in AI, a proposal for a regulation on artificial intelligence (AI), also known as the AI Act, to establish a risk-based and human-centric approach to developing, deploying, and using AI systems in the EU. In June 2021, representatives of the EU and their counterparts from the United States launched the Trade and Technology Council (TTC), a new forum to coordinate and cooperate on trade and technology issues and to lead the values-based global digital transformation.

In 2022, the EU adopted the Declaration on Digital Rights and Principles for the Digital Decade, which translates the EU's vision of digital transformation into principles and commitments for the protection and empowerment of citizens, the promotion of democracy and the rule of law, and the achievement of the EU's environmental and digital goals. The EU also enacted the EU Chips Act, a

comprehensive strategy to increase the EU's capacity and resilience in the design and production of semiconductors, which are essential for the digital transition. Furthermore, the EU adopted the EU Data Act, a legislative initiative to foster data sharing and reuse across the EU and to create a fair and competitive data economy. Finally, the DSA and the DMA entered into force in November 2022.

In June 2023, the European Parliament adopted its negotiating position for the AI Act, allowing the legislature to move to final, interinstitutional negotiations between the Commission, the Parliament, and the Council (also known as the Trilogue). The trilogue negotiations have faced several challenges and delays due to disagreements on three contentious areas: the use of remote biometric identification by law enforcement, the governance and enforcement of the regulation, and the regulation of foundation models or general-purpose AI systems.

Applying the Moral Landscape of AI Governance

The American and European approaches to AI have been studied from international relations and political science perspectives, showing their stark contrasts and points of alignment (Roberts et al., 2021; Burnay & Circiumaru, 2023; Crescenzi et al., 2007). In the following sections, I will propose an interdisciplinary framework utilizing the moral landscape as an alternative lens that studies policy in the richness of its context and with attention to the other actors in its complex socio-technical network (Iskandarova, 2017).

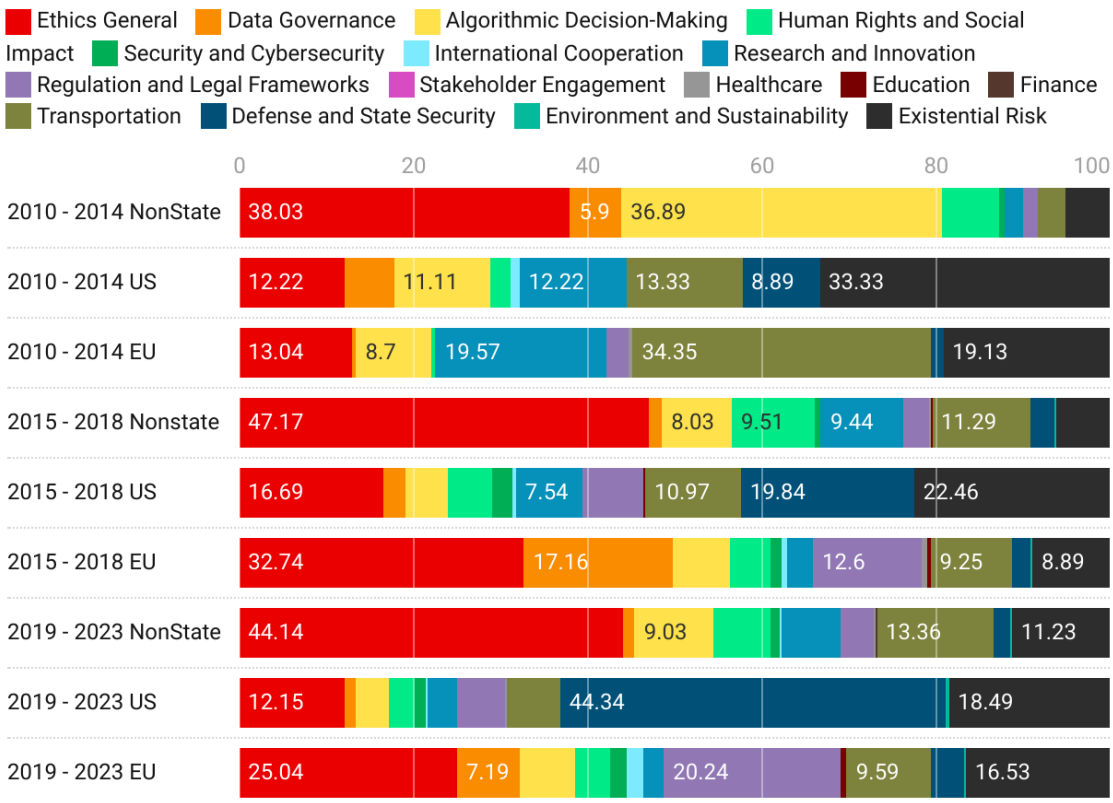

Figure 7. Theme occurrence in AI policy initiatives in the EU, the US, and the Non-state actors in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s (containing the period 2019 - 2023).

The moral landscape, consisting of a set of carefully defined themes, provides valuable insights into the fast-evolving nature of AI governance. This approach is beneficial for comparative study. The focus of this analysis is policy initiatives from two similar regions[14] represented in two corpora of policies addressing the same subject: artificial intelligence. The same international

[14] As Sartori noted, different entities are only comparable with respect to their shared traits (Sartori, 1991). The commonalities between the United States and the European Union were established in numerous comparative studies, for instance De Baere, G., and Gutman, K., (2012). 'Federalism and International Relations in the European Union and the United States: A Comparative Outlook'

corporations based in Silicon Valley developed this powerful, transformative technology.

The United States and the European Union represent two complex liberal democratic systems with multi-level governments (Hix & Høyland, 2022; Fabbrini, 2010; Brattberg & Rhinard, 2011). In the previous section, I introduced important historical, cultural, and economic differences in their regulatory logic. However, the EU and the US have an essential unifying factor in the fast adoption of AI technologies by their citizens, and the internal and external pressure to adopt a normative position to the development and deployment of these technologies. Both the United States and the European Union are economically and technologically advanced societies and have the same small set of large technology companies in entrenched and durable positions operating on their territory and serving their citizens (Bendiek & Stürzer, 2022; Slaughter & Cantwell, 2012; Coll-Mayor et al., 2007).

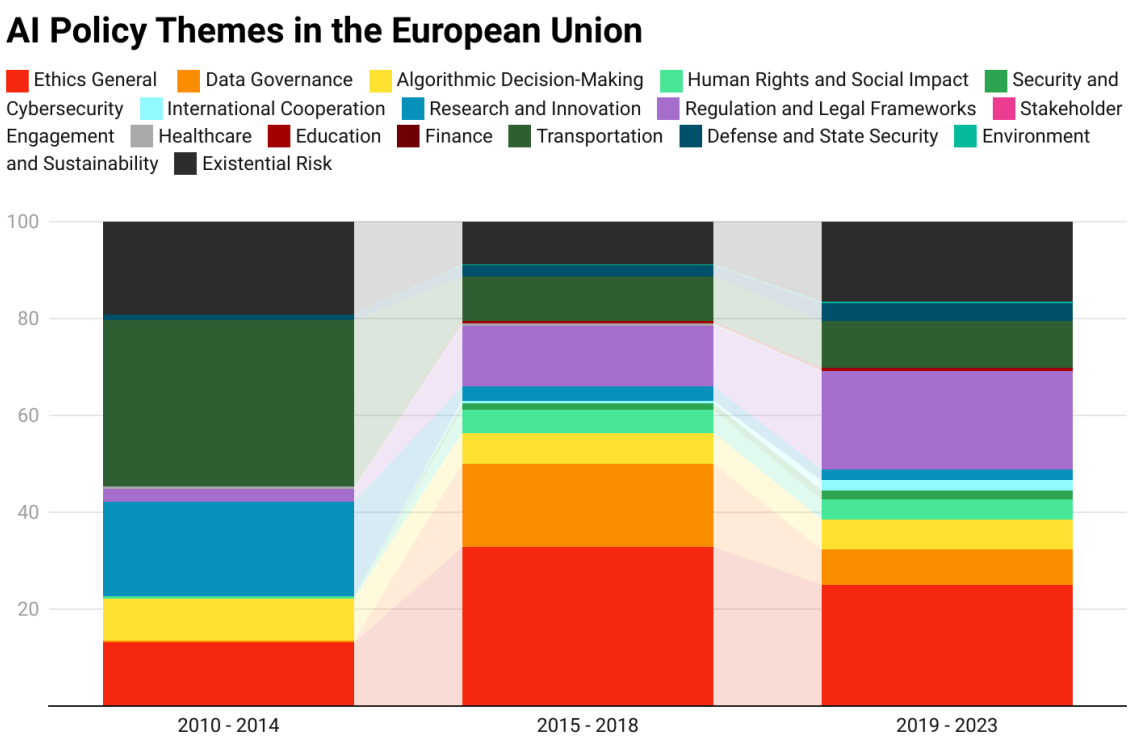

Figure 10. Theme occurrence in AI policy initiatives in the European Union in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s (containing the period 2019 - 2023).

Both the United States and the European Union have been devoting significant attention to the field of artificial intelligence (AI). This has led to the development of more policy initiatives in comparison to other regions worldwide.[15] It's worth noting that while the US has not opted for an AI-specific, legally binding intervention, the EU has taken a different approach by pursuing such measures for many aspects of AI. The EU justifies its stance by citing the protection of its own values that are challenged by the growing influence of

[15] To this date, the United States has published 82 AI Policy initiatives, and the European Union has published 63 AI Policy initiatives. Comparatively, China and Russia have published 22 and 11 initiatives related to AI, respectively (source: oecd.ai). Please refer to Appendices A and B of this dissertation for a list of AI policy initiatives in the two regions between 2010 and 2023.

foreign tech companies and the lack of solid and European-made alternatives. A couple of explanations may become evident from comparing the composition of the moral landscape themes on the two shores of the Atlantic. The European Union expressed the intent to position itself as a leader in championing its values globally, which is reflected in the growing focus on the AI ethics theme in its policy initiatives. However, compared to the US or non-state policy initiatives, its focus is significantly more diversified.

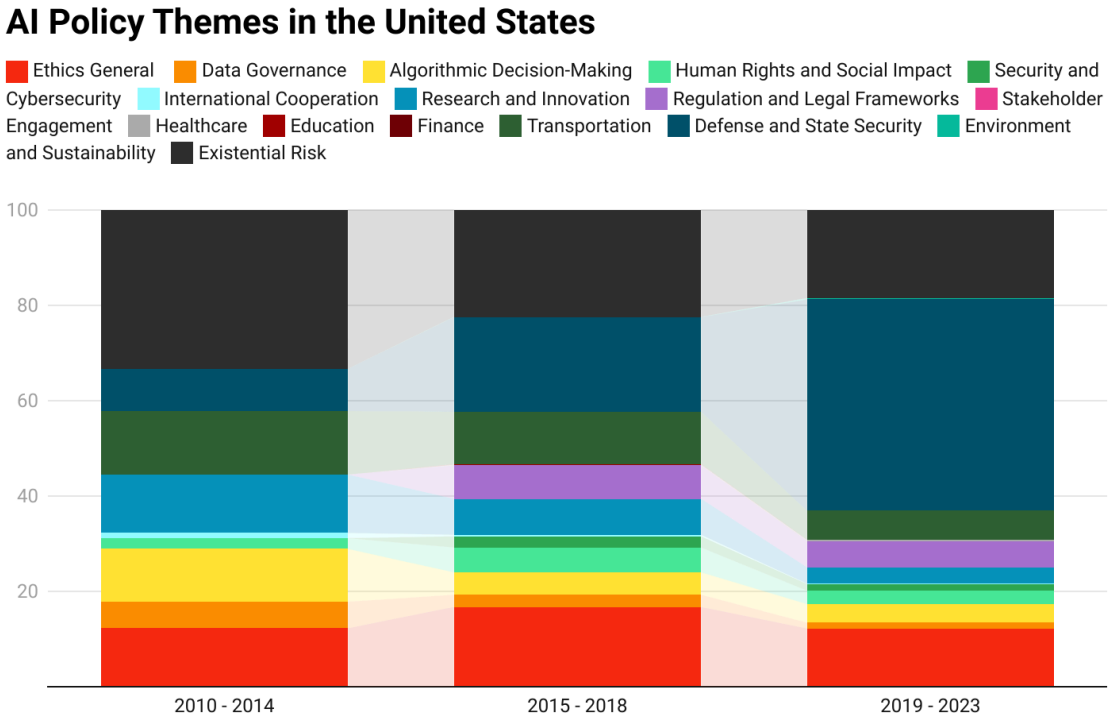

Figure 9. Theme occurrence in AI policy initiatives in the United States in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s (containing the period 2019 - 2023).

The EU's broad focus on diverse aspects of AI ties to its cultural and historical logic of anticipatory policymaking (Moser, 2023). Moreover, the

European Union does not maintain a military force where the United States prioritizes national security. In response to the Ex-Risk theme with its complex power dynamics explored in chapter two, embracing the anticipatory regulations filled the lack of military might. The US focus on defense and security reflects its most pressing challenge in maintaining its position of global dominance, which poses an existential threat to the nation. In contrast, the EU's priority to protect its citizens' rights and values makes it more responsive to the existential threat to humanity, motivating value promotion over focusing on defensive preparedness.

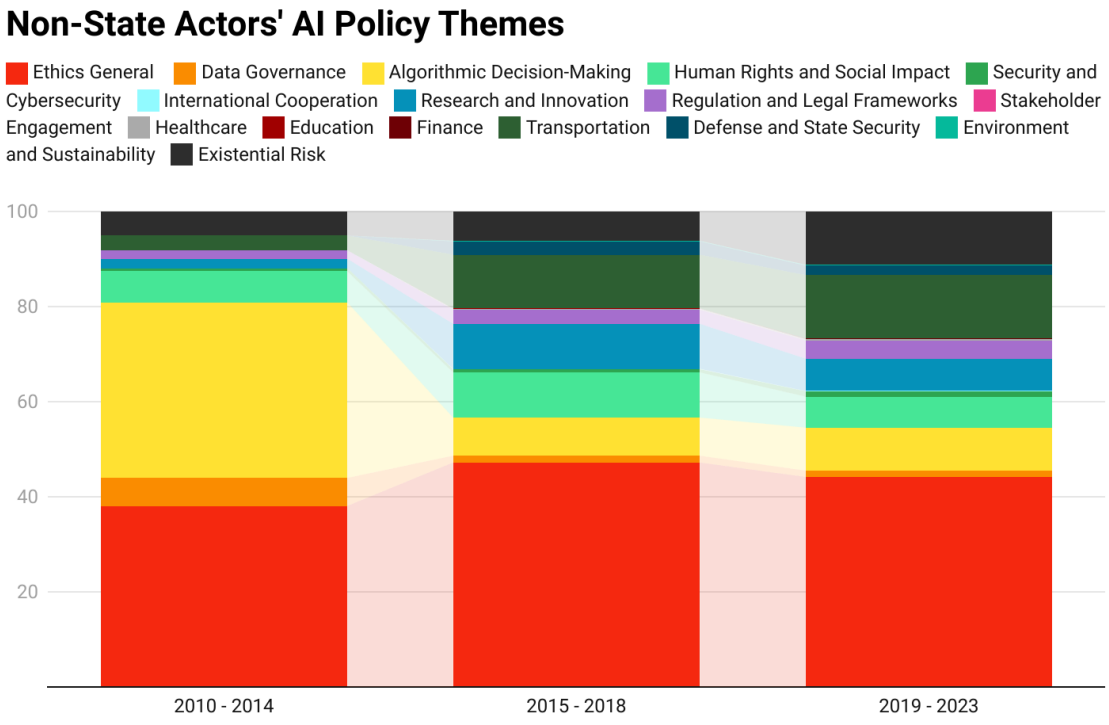

Figure 8. Theme occurrence in AI policy initiatives published by the Non-state actors in the early 2010s (containing the period 2010 - 2014), mid-to-late 2010s (containing the period 2015 - 2018), and Early 2020s

(containing the period 2019 - 2023).

The United States continues to engage in public-private partnerships, continuing its tradition of embracing innovation and minimal intervention. At the same time, the United States' position focuses more on defense and security, allowing the private sector to self-regulate. Meanwhile, the European Union enacts anticipatory policies and aims to lead discussions about the value of goods and services and the development of legal and regulatory frameworks. The non-state actors' continuous focus on ethics seems to be the most obvious choice, conveniently navigating the different regulatory environments and geopolitical tensions.

Chapter 4

Moving Forward: A Case Study in Strategic Public-Private Partnerships in AI Governance

Introduction

This thesis focused on understanding the moral landscape of AI and its key actors. Implementing AI policy initiatives, mainly those concerning AI ethics, remains a challenge. In particular, the private sector, which drives much of AI innovation, often falls short of translating ethical principles into practice (). We argue that the gap between principles and practice offers an opportunity for an intervention. This case study aims to address this gap by developing a software tool that assists practitioners in integrating AI ethics throughout the lifecycle of AI technologies. We put forward this case study as a contribution to the *third moment of AI ethics,* which we present as a direction for the AI ethics community writ large.

The AI Ethics community responded to the call for normative guidance of AI with a multitude of soft governance mechanisms, such as principles, frameworks, and guidelines (Morley, 2021; Hagendorff, 2022; Munn, 2023). These documents were critical for the development of a shared normative foundation about AI, revolving around core principles modeled after the classic

Medical Ethics Principles. It is now well-established that AI technologies should be beneficial to people and the environment (beneficence); robust and secure (non-maleficence); respectful of human values (autonomy); fair (justice); and explainable and accountable (explainability) (Jobin, 2019; Floridi, 2022; Schiff 2020; Morley, 2020; Martinho, 2022; Mittelstadt, 2019).

On a second moment, the AI Ethics community focused on operationalizing these principles and guidelines (Prem, 2023; Ashok, 2022; Morley, 2021; Morley, 2020; John, 2022; Ayling, 2022; Vakkuri, 2021; Stix, 2021; Wilson, 2022; Georgieva, 2022; Johnson, 2021; Eitel, 2021; Theodorou, 2020). The Morley Typology, a staple in this so-called *from what to how* moment of AI Ethics, is a comprehensive framework that matches available methods and tools to the core AI Ethics principles (*Beneficence, Non-Maleficence, Justice, Autonomy, and Explainability*) and to stages in the algorithm development pipeline (*Business Development; Design; Training and Test; Prototyping; Testing; Development; Monitoring*) (Morley, 2020).

However, it seems that these principles and tools have yet to be integrated into technology R&D practices (Hagendorff, 2022). Several reports indicate their minimal impact on professional practices (Vakkuri, 2020; Orr, 2020; Johnson, 2021; Vakkuri, 2022; Munn, 2023). One factor contributing to this challenge is that these normative mechanisms are presented as external entities concerning particular science and technology domains, thus lacking *relatability* (Morley, 2021).

We consider that, as AI continues to become more embedded in different aspects of society, the AI Ethics community should focus on improving the relatability and context specificity of the normative tools. We are aligned with scholars who have called for AI Ethics to develop tools that go beyond general professional and ethical practice guidelines, reflect normative diversity, do not require a deep background in philosophy, and ensure all stakeholders are engaged and involved in design decisions (Morley, 2021).

In this research, we first reflect on the challenges of AI Ethics and subsequently describe an AI Ethics tool that we developed in what we propose to be the third moment of AI ethics. The tool is grounded in the current normative literature of AI Ethics, yet it is also practical. We included a case study in the Autonomous Driving industry and requested input from practitioners working in autonomous driving startups in the United States.

The Challenges of AI Ethics

AI Ethics as a Branch of Applied Ethics

Ethics explores matters of right and wrong, reflecting the moral spectrum of a particular concept, policy, or technology (Martinho, 2022). The normative debates welcome disagreement, speculation, and abstraction, thus allowing both diversity of thought and serious ethical consideration (Mittelstadt, 2019; Martinho, 2022). These debates slowly build a robust, rich, and diverse normative foundation, which eventually serves as the basis for governance and regulatory mechanisms.

AI Ethics explores the ethical implications of AI technologies. It scrutinizes the moral issues arising from the design, development, deployment, and use of AI systems. Like other branches of Applied Ethics, such as Medical Ethics, normative principles and tools may be used to reflect on complex moral dilemmas, but they also aim to guide practice (i.e., technology development and innovation).

The community working in this space has focused mainly on translating and incorporating normative elements through the design pipeline. These issues are deeply embedded in the larger societal context, namely in domains such as Transportation (Martinho, 2021), Healthcare (Martinho, 2021), and Justice (Martinho, 2024). The socio-technical nature of AI prompts ethicists to examine the role of software developers and other practitioners in the technology's outcomes.

The Challenges of AI Ethics in the Current AI Paradigm

AI Ethics faces unique challenges related to the urgency of normative guidance, the multi-purpose nature of AI, and a multitude of stakeholders (Martinho, 2022). The recent state of affairs in AI is characterized by the fast development and deployment of these technologies, as well as an urgent need for practical and operational normative guidance, resulting in an *AI Ethics Boom* (Correa, 2023), where many academic, government, and industry organizations started publishing normative guidance to AI. These efforts were predominantly concerned with the principles that ought to guide the *modus operandi* of

practitioners; however, as mentioned earlier, there needed to be more guidance on implementing such principles into practice (Correa, 2023).

Another challenge for AI Ethics is related to the fact that modern AI is a multi-purpose technology. AI is currently applied to several societal domains, such as Transportation, Healthcare, and Justice (Martinho, 2022). AI Ethics needs to consider how risks, conflicting rights and interests, and social preferences vary in different contexts (Morley, 2021); like in other Applied Ethics domains, namely Medical or Business Ethics, many distinct issues require rigorous examination and scrutiny.

Due to the multi-purpose nature of AI, there are many stakeholders involved in developing various AI-powered technologies. Organizations such as Google, Amazon, Facebook, and OpenAI have shown a strong interest in AI Ethics and have developed several guidelines (Hagendorff, 2020; Ali, 2023). Traditionally, the Ethics community is hesitant about industry-led initiatives regarding Ethics and often dismisses these initiatives as attempts to shape the normative conversations to their interests (Morley, 2020; Jobin, 2019).

However, it may also be argued that these organizations produce guidelines in response to the Ethics communities (as well as concerns and criticism in popular culture and media). The Ethics work is rarely *ready to use* as it is often abstract, speculative, and intricate.

A challenge for AI Ethics researchers is to operationalize their work and communicate effectively with other communities in the AI space (Martinho,

2022). Experience with other applied fields of Ethics (e.g., Medical Ethics) shows that it is possible to operationalize Ethics successfully from abstract ethical principles, thus lending credence to the AI Ethics effort (Morley, 2021). This research builds on the premise that AI Ethics needs to focus on improving the relatability and contextualization of its normative tools (Martinho, 2022).

The Three Moments of AI Ethics

The AI Ethics work developed in recent years meets traditional expectations in terms of richness, controversy, and diversity of thought (Martinho, 2022). There have been numerous contributions, including guidelines, white papers, or company statements of commitment to ethical design (High-Level Expert Group on AI 2019; Hilligoss & Fjeld, 2019; European Commission 2020). Still, the research on the preparedness of practitioners to face ethical dilemmas in the design and development of artificial intelligence implies a disconnect between established ethical norms and practice (Brightman, 2018). To better understand the disconnect between research and practice, we revisit the three moments in the field of AI ethics.

In the first moment of AI Ethics, as mentioned earlier, a multitude of soft governance mechanisms were developed by the scientific community, governments, private companies, and non-governmental organizations (Jobin, 2019; Hagendorff, 2020). The subject of ethical inquiry was the notion of alignment between moral values and the machine objectives (Gabriel, 2020), and much of the research efforts focused on identifying the principles and values that

ought to guide the software engineering practice (Renda, 2019; Floridi, 2018). However, ethics principles have been deemed abstract, not relatable, and yet to be integrated into technology practices, sparking concerns about the success of the AI Ethics endeavor.

Moreover, there is an ongoing mistrust that some organizations that contributed to this normative corpus, namely private organizations involved in the development of AI technologies, will not go beyond *ethics washing* and implement ethical practices voluntarily. When Ethics is integrated into organizations, there are concerns that it is used merely as a marketing strategy with little impact when it comes to decisions (Hagendorff, 2020; Orr, 2020). A study that surveyed AI practitioners about their perceived impact of AI Ethics guidelines reported that the effectiveness of such guidelines or ethical codes is extremely low and that they do not change the behavior of professionals from the tech community (Johnson, 2021).

In a second moment, the AI Ethics community focused on translating abstract principles to more specific practice-oriented guidelines (the so-called \textit{from what to how} phase in AI Ethics (Morley, 2020). However, there is little evidence of the impact of these tools and methods on the governance of AI (Morley, 2020).

The research shows that ethical guidelines do not translate into practice. Studies have reported that practitioners do not see ethics as a concern or challenge during the development of Machine Learning models and do not account for ethics in their practices, even though they might find ethical guidelines overall

useful for their organizations or care about ethics on a personal level (Vakkuri, 2020; Johnson, 2021; Munn, 2023).

The research shows that there needs to be more ethics and technology research and its application in engineering (Karim, 2017). Engineers perceive ethics as a *system of barriers and constraints,* and the complex language of ethics prevents its application and deployment throughout the innovation process (Byers, 2021).

As mentioned earlier, Morley et al. identified the tools and methods already available to guide AI practitioners on core issues of AI Ethics and plotted these methods and tools in a typology, matching them to ethical principles and stages in the algorithm development pipeline. They reported that numerous tools and methodologies exist to assist practitioners in realizing Ethical AI, but the vast majority are severely limited in terms of usability (Morley, 2020).

In the third moment of AI Ethics, the community needs to focus on developing tools that are relatable and contextualized. We align with Morley et al.: AI Ethics requires an approach that (i) goes beyond general guidelines for professional and ethical practice, (ii) embodies a toolset that does not require a deep background in philosophy, (iii) reflects the normative status of ethical reasoning, (iv) ensures all stakeholders are engaged and involved in design decisions rather than simply consulted about them; and (v) reflects the language and tensions in particular fields (Morley, 2021).

In this research, we contribute to the third moment of AI Ethics. We built on the Morley Typology and developed a normative tool that is theoretically grounded, diverse, and practical and can be contextualized in different domains. By translating abstract moral values such as fairness or transparency into specific applicable items that can help guide technological development, we provide a tool and language that resonates with the practitioners and empowers them to engage in complex conversations over the socio-technical nature of their work.

The AI Ethics Tool

Review of the Morley Typology

The AI Ethics tool is inspired by the Morley Typology (Morley, 2020). This theoretical framework matches N = 106 available methods and tools reported in the literature to the core AI Ethics principles (Beneficence, Non-Maleficence, Justice, Autonomy, and Explainability) and to phases in the algorithm development pipeline (Business Development; Design; Training, and Test; Prototyping; Testing; Development; Monitoring). We first reviewed the literature associated with each phase and excluded the methods and tools no longer available. We used this compressed Morley Typology as our starting point for the development of the AI Ethics tool, but we also reviewed and included methods and tools beyond the Typology that were also aligned with such mandates. We believe that the current version of the Tool is representative of the diverse normative work in the AI Ethics literature.

Software Development

The tool was developed in Gitbook (https://ai-ethics-tool.gitbook.io/ai-ethics-tool), a platform popular for capturing and documenting technical knowledge, such as product documentation, internal knowledge bases, or Application Programming Interfaces. Gitbook also enables Github integration, which allowed us to build our tool as an open-source project. The tool is divided into sections reflecting the algorithm development pipeline, adjusted from the Morley typology (Morley, 2020) as (i) Development, (ii) Design, (iii) Training, (iv) Building, (v) Testing, (vi) Deployment, (vii) Monitoring and (viii) Fostering Ethics and Virtues. Our technology-agnostic, open-source model under the MIT license allows practitioners to create their own version suitable for their specific needs. As an example of further operationalization, we developed a tab specifically aimed at the autonomous vehicles industry. To further improve the intuitive interface, we implemented an AI-powered semantic search, which allows users to search in natural language and returns answers summarizing the content and linking directly to relevant information within the website.

Testing the Tool

Case Study on Autonomous Driving

We developed an empirical study focusing on the Autonomous Driving industry to test the tool and assess its relatability. We consider that this industry makes a compelling case study, as the Autonomous Vehicle (AV) has been widely

used in the AI Ethics literature to illustrate the ethical trade-offs in traffic situations that entail the distribution of risk (Bonnefon, 2016; Bonnefon, 2019; Martinho, 2021). The particular aims of the case study are (i) to engage practitioners in the AV community; (ii) to assess the relatability of the tool; (iii) to receive feedback and make improvements in the tool; and (iv) to follow up on previous studies and investigate whether this approach has the potential to be useful for practitioners.

Survey

We developed a survey using Qualtrics (https://www.qualtrics.com.), a software for creating and implementing online surveys. The study was submitted to the Tufts University Institutional Review Board (IRB) and received an exempt determination. The final version of the survey features thirteen groups of questions (Supplementary Materials). The first group of questions relates to (i) Professional Role & Experience; (ii) Knowledge of Ethics; (iii) Perspectives about AI Ethics; (iv) AI Ethics in Industry Projects; (iv) Relatability, Usefulness, and Feedback on the Tool.

Results

We deployed the survey via email to (N=40) Autonomous Driving companies in January 2024 (Supplementary Materials). The engagement in the survey was low, and we recorded only nine responses (seven with no missing data). The data was stored in Qualtrics. The results shown below relate to the characterization of the study participants, their (self-perceived) knowledge and

perspectives about (AI) Ethics in general (Penuel, 2017) and in their projects, and also the reliability and usefulness of the tool (Tables 1-5). The comments and feedback provided by the participants are included in Appendix D.

Professional Role	**Participants (N)**
Software Development	5
Product Management	1
R&D	3
Experience	
Entry Level (0-3 years)	2
Intermediate (3-7 years)	3
Mid-Level (8-15 years)	3
Senior-level (15+ years)	1

Table 3. Professional Role and Experience of Participants

Ethics	62
AI Ethics	60
AV Ethics	63

Table 4. (Self-Perceived) Mean Knowledge of Ethics [0-100]

Phase	Participants
Development	
Sometimes	5
About half the time	3
Most of the time	1
Design	
Sometimes	3
About half the time	3
Most of the time	1
Always	2
Training	
Never	2
Sometimes	1
About half the time	3
Most of the time	2
Always	1
Building	
Never	1
Sometimes	1
About half the time	5
Most of the time	2
Testing	
Sometimes	2
About half the time	2
Most of the time	4
Always	1
Deployment	
Sometimes	3
About half the time	1
Most of the time	4
Always	1
Monitoring	
Sometimes	1
About half the time	2
Most of the time	4
Always	2

Table 5. AI Ethics in Industry Projects

Statement	Participants
AI Ethics addresses questions that help my team make better decisions	
Somewhat agree	6
Neither agree nor disagree	3
AI Ethics researchers live in an ivory tower isolated from practice	
Somewhat agree	4
Neither agree nor disagree	5
The AI Ethics claims are trustworthy	
Somewhat agree	4
Neither agree nor disagree	5

Table 6. Perspectives About AI Ethics

Feedback	Participants
Relatability	
Relatable	9
Adoption	
Yes	3
No	3
Flow	
Yes	7
No	2

Table 5: Relatability, Usefullness, and Feedback on the Tool

Table 7. Relatability, Usefulness, and Feedback on the Tool

Discussion & Conclusions

As a novel Applied Ethics field, AI Ethics has experienced rapid development, reflecting the emerging need for normative discussion around this transformative technology. Despite the richness of AI Ethics research, there are concerns about its success in realizing normative endeavors, especially in light of the latest developments in Generative AI (Hagendorff, 2024). Empirical research shows that even when ethical guidelines resonate with developers, these normative efforts do not translate into practices (Vakkuri, 2020; Johnson, 2021; Munn, 2023; Johnson, 2021; Byers, 2021).

We contend that AI Ethics needs to move to its Third Moment, in which it is contextualized within particular scientific and technology domains, to reflect the language, tensions, and priorities of such domains and, therefore, improve relatability. Understanding and utilizing practices and tools that practitioners use further enhances their integration into the operational processes of a specific team or organization.

We developed a theoretically grounded yet practical AI Ethics tool that combines numerous translational tools and methods to assist practitioners in incorporating Ethics into their practices. We included a domain-specific tab related to Autonomous Driving. We also developed a small empirical study to test this tool by surveying practitioners currently working in the Autonomous Driving space. Although this study has clear limitations related to the small sample of practitioners who participated in the study (N=9), we received a good amount of feedback that allowed us to further improve the tool.

Our results are aligned with previous empirical studies that indicate that practitioners acknowledge the importance of AI Ethics but do not account for it in their practices (Vakkuri, 2020; Johnson, 2021; Munn, 2023). The participants in our empirical study indicate they have some knowledge about Ethics, AI Ethics, and Autonomous Driving Ethics but are somewhat neutral about their perspectives on AI Ethics and AI Ethics researchers. They also indicated that Ethics comes up more often in the later phases of the innovation pipeline. Regarding our tool, the participants considered that it is relatable and the flow of the tool is adequate. The participants also provided extensive written feedback regarding potential barriers to adoption that will be incorporated into the next iteration of the tool.

Through the analysis of the comments, it is clear that the practitioners who provided feedback on the tool still wish for more detail, precise instructions, and examples. Participants often mention that the language is still high level, thus reinforcing once again the need for relatable and contextualized AI Ethics tools.

As AI Ethics moves forward, the community needs to reflect on the best ways to improve the relatability and context-specificity of the normative tools, namely through collaborative endeavors, so that different communities can further engage and contextualize AI Ethics. The private sector has a dominant position in much of the research and development of AI. Industry self-regulation can help to ensure that AI systems are developed and used in a way that aligns with the values and interests of society. However, self-regulation may be misinformed or lack crucial perspectives to address the potential risks and impacts of AI, especially in cases where the technology is used in areas such as healthcare or public safety. Research-grounded interventions such as the introduced AI Ethics Tool provide a more accessible venue for practitioners to meaningfully engage in conversations around ethics and the normative aspects of their work while utilizing existing processes and reflecting familiar practices.

Chapter 5

Conclusion

"Many of our most serious conflicts are conflicts within ourselves. Those who suppose their judgements are always consistent are unreflective or dogmatic."

- John Rawls, Justice as Fairness

Reflecting the Role of Non-State Actors in the Governance of Large Language Models

Examples of early regulating efforts in emerging powerful AI applications help illustrate non-state actors' role in AI governance. This concluding chapter reflects on large language models (LLMs), which have been a blooming research domain with applications spanning from language translators to text editor tools to smart assistants or predictive text and search results (Brown et al., 2020; Jurafsky & Martin, 2002; Nadkarni, Ohno-Machado, & Chapman, 2011).

Enormous data sets and scalability allowed for large language models to shift the paradigm around automated natural language processing. The difference between previous approaches in NLP (Natural Language Processing) and LLMs is not solely in the amount of data used. The approach from previous sequential processing of the text to a more sophisticated, deeper analysis of how words in sentences are related to each other (Radford et al., 2019; Tamkin et al., 2021). Combined with extremely large datasets, these models can perform various tasks for which they were not formally trained. This is possible due to the change in

approach in the model's training, where the model recognizes and retains complex language patterns. With the growing size of training data, these models' accuracy was proven to improve so significantly that they can now perform correctly in novel situations (Okerlund, et al., 2022).

Large language models (LLMs) have the potential to be extremely useful, but there are also several drawbacks to their use, such as the environmental impact, the reinforcement of discrimination or marginalization of certain groups or individuals, or costs that restrict their use to very elite, established organizations or research groups. Training a single LLM was estimated to emit comparable amounts of CO2 to a transatlantic flight. One training period usually spans multiple days and requires additional computational procedures for increased accuracy in a specific task. For example, the BERT model developed by Google is estimated to produce 284 tons of CO2 during training, which is equivalent to the annual carbon emissions of 50 people (Bender, Gebru, & et. Al., 2021; Strubell et.al., 2019).

The challenges of LLMs are not limited to their environmental impact. Potentially biased outcomes can be traced back to the source of their training data. Many models have been found to reflect the biases and ideologies of the Western, English-speaking world and the overrepresentation of certain groups. For example, the GPT-3 model developed by OpenAI was found to exhibit bias against Muslims by frequently suggesting violent actions when completing sentences involving this group (Abid, Farooqi, & Zou, 2021).

Regulatory oversight may be needed to address the challenges posed by LLMs (and other applications of artificial intelligence research), which have already had significant impacts and are likely to continue. In the earlier chapters, I show that these challenges, including environmental impacts and biases in machine learning, are being discussed and addressed by both government institutions and non-state actors. The rapid advancement of artificial intelligence technologies has made it challenging for regulators to provide adequate oversight and ensure that appropriate measures are in place. Ongoing dialogue and collaboration between government, industry, and other stakeholders seemed necessary to address this. The balance of power between state and non-state actors plays a significant role in shaping the direction and impact of AI. Artificial intelligence governance's success may depend on all actors' ability to work together to anticipate and address the challenges and opportunities that artificial intelligence presents.

As I discuss in Chapter One, cutting-edge LLMs are developed by private entities with international end users. Some of the most impactful large language models, known for their innovative design and large training data size, have been developed by international corporations. Some notable examples are Google's BERT and OpenAI's GPT-4. However, another notable, popular text-to-image model, Stable Diffusion, was developed in an open-source environment by Stability AI (Okerlund, et al., 2022).

OpenAI's management of the GPT-4 model illustrates how a dominant player in the artificial intelligence (AI) industry can shape the conversation and

priorities related to AI governance. The company keeps developing models using extremely large training data sets, which have been criticized for their environmental impact and unsustainable practices. At the same time, the company has tried to address bias in the model and avoid harmful outputs. Findings of their risk assessments were published with a response from the company to specific fixes to found problems. These self-regulatory efforts also included filtering the use cases to prevent the model from creating harmful content and developing an extensive veto process for new user accounts.

The priorities of a dominant player in the AI industry can influence the broader conversation and direction of AI governance. In contrast to OpenAI's approach to the governance of AI, the much smaller startup company, Stability AI, has released its Stable Diffusion text-to-image model as open source with no barriers or filters to users' creativity. Consequently, this model has been reported to generate harmful imagery of violence and pornography featuring real individuals (Milne, 2023). In response to criticism, the CEO of Stability AI, Emad Mostaque, has emphasized the company's commitment to equal access to technology and freedom of expression as priorities. The two companies landed on opposing positions regarding the potential trade-offs and challenges involved in different approaches to AI governance, including the balance between access and oversight. Stable Diffusion’s official release announcement includes a caveat regarding potentially biased results given the nature of the training data, along with an ambiguous invitation to open dialogue regarding these limitations to their

technology. The company encourages ethical use but stresses individual responsibility (STABILITY AI LTD, 2022).

United States congresswoman Anna G. Eshoo issued a press release shortly after the Stability AI released its model, calling for an intervention from the National Security Advisor (NSA) and the Office of Science and Technology Policy (OSTP). Citing the potential societal harm of the technology, Eshoo specifically compared the Stable Diffusion model to OpenAI's GPT-3 based text-to-image model Dall.E. Eshoo cites democratizing access to AI as her objective in her legislative work. She condemns, however, the open-source release of the Stable Diffusion model as careless and unsafe. Eshoo highlights that Open AI's model blocks harmful outcomes such as propaganda, misinformation, or explicit content (Eshoo, 2022).

Eshoo's call for intervention from the National Security Advisor and the Office of Science and Technology Policy reflects the concern that the potential societal harm of the Stable Diffusion model may require a more coordinated response from the US government. While the US has taken a more laissez-faire approach to regulating AI, there is recognition that certain applications of AI, such as those with the potential to cause harm, may require more stringent oversight and regulation. Eshoo used the comparison of the Stable Diffusion model to OpenAI's Dall.E to highlight the need to ensure that AI is developed and deployed responsibly and ethically and that the risks and impacts of AI are carefully considered.

Eshoo's call for intervention also highlights non-state actors' role in AI governance. Tech companies and other non-state actors have taken the lead in the development and deployment of AI, and their decisions and practices significantly shape the direction and impact of AI. It is important to keep critically examining the balance of power between state and non-state actors in AI governance and the need for democratic oversight to ensure that AI is developed and used in a way that aligns with the values and interests of society. The congresswoman sided with the dominant player, further strengthening its position while forcing the smaller competitor to "fall in the line," which would likely cause Stability AI to lose some of the advantages inherent to their different business model and company ethos.

Soon, it is likely that more regulators will engage with large language models as these models become increasingly prevalent and influential in society. The AI Act in the European Union received an additional section grappling with LLMs, as their general-purpose nature was not reflected in the earlier regulation draft (Yasar et al., 2024). Large language models can significantly impact various sectors and industries, including finance, healthcare, education, and media. As such, it is important for regulators and, even more so, for the general public to understand the risks and opportunities associated with these models. The involvement of both is a first step to ensure they are used responsibly and ethically (Solaiman, 2019).

For example, LLMs may perpetuate biases and stereotypes or be used to spread misinformation or propaganda (Leavy, 2018; Perez, 2019; Zuboff, 2019). Tech companies have already started to develop best practices for their use. It is

crucial, however, that many more relevant stakeholders get a seat at this table. As large language models become more prominent, the need for regulators and civil society's much-needed input in their development and use will only increase (Palladino, 2021).

Discussion & Future Directions for AI Governance Research

This dissertation focused on the unique position and power dynamics between non-state actors and the US and EU regulatory regimes in AI governance. Using a mixed-method, interdisciplinary approach, it presents a contribution to the field that provides evidence and resources for future work in data analysis and a new framework of the moral landscape of AI governance. The moral landscape framework combined quantitative approaches using the Moral Landscape Dictionary and qualitative methodology based on the actor-network theory.

Tensions between seemingly opposing camps define the field of AI governance: the AI ethicists who argue we should focus more on social justice issues in contemporary AI applications and long-term thinking technologists who argue for more preparedness in case catastrophic scenarios become a reality. There are disciplinary divides and mistrust between practitioners and regulators. This dissertation contributes to AI governance by clarifying the complexity of these prominent issues in the field and providing more nuanced explanations.

The presented results aim to empower practitioners, regulators, and civil society organizations by providing data in the context of historical developments and sociopolitical evolution of AI governance as both a research field and a practice. This dissertation presents an alternative interdisciplinary approach to AI governance research. It contextualizes technical knowledge in the rich, interconnected network of institutions, individuals, and artifacts and studies their complex relationships. This approach challenges intuitions about AI that disenfranchise the general public and civil society in Western democracies from expressing their vision for AI as a public good.

An important future direction for this research is hence to expand the group of non-state actors to civil society organizations and advocacy groups, whose voices are too often underrepresented in the mainstream media and public discourse but who are emerging to take on the much-needed work (Floridi & Cowls, 2022; Bahçecik, 2019).

Bibliography

1. Dennett, D. C. (2007). *Breaking the spell: Religion as a Natural Phenomenon*. Penguin UK.
2. Arendt, H. (1958). The Human Condition (2nd ed.). University Of Chicago Press.
3. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Retrieved from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf
4. Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. New York, NY: Picador.
5. Jasanoff, S. (2016). The Ethics of Invention: Technology and the Human Future. W. W. Norton.
6. Wijninga, P., Oosterveld, W. T., Galdiga, J. H., & Marten, P. (2014). State and Non-State Actors: Beyond the Dichotomy. In J. van Esch, F. Bekkers, S. De Spiegeleire, T. Sweijs (Eds.), Strategic Monitor 2014: Four Strategic Challenges. The Hague: Hague Centre for Strategic Studies. Retrieved from https://www.jstor.org/stable/resrep12608.8
7. Maurer, S. M. (2017). Self-governance in science: Community-based strategies for managing dangerous knowledge. Cambridge University Press.

8. Josselin D. & Wallace W. (2001). Non-state actors in world politics. Palgrave. https://doi.org/10.1057/9781403900906
9. Leung, J. 2019. Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies [PhD thesis]. University of Oxford.
10. Ding, J., & Dafoe, A. (2021). The Logic of Strategic Assets: From Oil to AI. Security Studies, 1-31. Retrieved from https://arxiv.org/pdf/2001.03246.pdf
11. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Askell, A., Child, R., Dhariwal, P., Agarwal, S., Ramesh, A., Neelakantan, A., Herbert-Voss, A., & Ziegler, D. M. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165v4.
12. Jurafsky, D., & Martin, J. H. (2002). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education Asia. ISBN 81-7808-594-1.
13. Dowding, K. (2008). Agency and structure: Interpreting power relationships. Journal of Power, 1(1), 21–36. https://doi.org/10.1080/17540290801943380
14. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
16. Tamkin, A., Singh, T., Giovanardi, D., & Goodman, N. (2020). Investigating Transferability in Pretrained Language Models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1393-1401). Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.125
17. Arnold, Z., Rahkovsky, I., & Huang, T. (2020). Tracking AI Investment: Initial Findings from the Private Markets. (Center for Security and Emerging Technology, September 2020). doi:10.51593/20190011
18. Okerlund, J., Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M., & Parthasarathy, S. (2022). What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them. [Report]. Retrieved from https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox
19. Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 610-623). Association for Computing Machinery. doi:10.1145/3442188.3445922

20. OECD.AI Policy Observatory, (2022). OECD AI Principles overview. Retrieved from https://oecd.ai/en/ai-principles
21. OECD. (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Retrieved from https://oecd.ai/en/ai-principles
22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics. Retrieved from [https://aclanthology.org/P19-1355/]
23. Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. Nature Machine Intelligence, 3(6), 461-463. doi:10.1038/s42256-021-00359-2
24. Prunkl, C. E. A., Ashurst, C., Anderljung, M., et al. (2021). Institutionalizing ethics in AI through broader impact requirements. Nature Machine Intelligence, 3(2), 104-110. doi:10.1038/s42256-021-00298-y
25. Prunkl, C., & Whittlestone, J. (2020). Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. Retrieved from https://doi.org/10.48550/arXiv.2001.04335
26. Cave, S., & ÓhÉigeartaigh, S. S. (2019). Bridging near- and long-term concerns about AI. Nature Machine Intelligence, 1(1), 5-6. doi:10.1038/s42256-018-0003-2

27. Eshoo. (2022). Press release: Eshoo Urges NSA & OSTP to Address Unsafe AI Practices. Retrieved from [https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-unsafe-ai-practices]
28. Federal Trade Commission Act of 1914 (Law 15 U.S.C. §§ 41-58, as amended). (n.d.). Retrieved from http://uscode.house.gov/view.xhtml
29. Health Insurance Portability and Accountability Act (HIPAA). (1996). Pub. L. No. 104-191, 110 Stat. 1936.
30. National Institute of Standards and Technology (NIST). (2020). NIST AI framework: technical report (NISTIR 8266). Gaithersburg, MD: NIST.
31. Bietti, Elettra. 2020. "From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 210-19.
32. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203. https://doi.org/10.48550/arXiv.1908.09203
33. Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st International Workshop on Gender Quality in Software Engineering (pp. 14-16).

34. Perez, C. C. (2019). Invisible women: Exposing data bias in a world designed for men. London, UK: Chatto & Windus.
35. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. London, UK: Profile Books Ltd.
36. Palladino. (2021). The role of epistemic communities in the "constitutionalization" of internet governance: The example of the European Commission High-Level Expert Group on Artificial Intelligence. Telecommunications Policy, 45(6), 102149–. https://doi.org/10.1016/j.telpol.2021.102149
37. GDELT Project. (n.d.). Data: Querying, Analyzing, and Downloading. Retrieved October 22, 2023, from https://www.gdeltproject.org/data.html
38. Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, Hui Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT", Discrete Dynamics in Nature and Society, vol. 2017, Article ID 8180272, 13 pages, 2017. https://doi.org/10.1155/2017/8180272
39. Murali, R., Patnaik, S., & Cranefield, S. (2021). Mining International Political Norms from the GDELT Database. In A. Aler Tubella, S. Cranefield, C. Frantz, F. Meneguzzi, & W. Vasconcelos (Eds.), Coordination, Organizations, Institutions, Norms, and Ethics for

Governance of Multi-Agent Systems XIII (pp. 33-44). Springer International Publishing. https://doi.org/10.1007/978-3-030-72376-7_3

40. Roxana Radu, Steering the governance of artificial intelligence: national strategies in perspective, Policy and Society, Volume 40, Issue 2, June 2021, Pages 178–193, https://doi.org/10.1080/14494035.2021.1929728
41. OECD.AI (2021), powered by EC/OECD (2021), a database of national AI policies, accessed on 23/10/2023, https://oecd.ai.
42. D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.
43. Palladino, N. (2020). The role of the epistemic communities in the ' constitutionalization ' of the Internet Governance : the case of the EU High-Level Expert Group on Artificial Intelligence.
44. Latour, B. (1999). On recalling ANT. The sociological review, 47(1_suppl), 15-25.
45. Latour, B. (2005). Reassembling the Social: an Introduction to Actor-network-theory. Oxford University Press.
46. Latour, B. (2004). Politics of nature: How to bring the sciences into democracy (C. Porter, Trans.). Cambridge: Harvard University Press.
47. Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. Paper presented at the Proceedings of the 52nd Hawaii International Conference on System Sciences.

48. Stigler, G. J. (1971). The Theory of Economic Regulation. The Bell Journal of Economics and Management Science, 2(1), 3–21. https://doi.org/10.2307/3003160
49. National Artificial Intelligence Advisory Committee. (n.d.). Home. Retrieved November 1, 2023, from https://ai.gov/naiac/
50. Office of Science and Technology Policy. (2022, February 3). OSTP's continuing work on AI technology and uses that can benefit us all. The White House. https://www.whitehouse.gov/ostp/news-updates/2022/02/03/ostps-continuing-work-on-ai-technology-and-uses-that-can-benefit-us-all/
51. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST.AI.100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
52. Dal Bó, Ernesto (2006). "Regulatory capture: A review". Oxford Review of Economic Policy. 22 (2): 203–225. doi:10.1093/oxrep/grj013. JSTOR 23606888.
53. Amba Kak and Sarah Myers West. (2023). AI Now 2023 Landscape: Confronting Tech Power. In: AI Now Institute, April 11, 2023. Retrieved from: https://ainowinstitute.org/2023-landscape.
54. European Commission. (2023). Digital Markets Act: Commission designates six gatekeepers. Retrieved from: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4328

55. The White House. (2016a). Preparing for the future of artificial intelligence.
56. The White House. (2016b). The national artificial intelligence research and development strategic plan. Retrieved from: The National Artificial Intelligence Research and Development Strategic Plan (archives.gov)
57. The White House. (2019, February 11). Accelerating America's leadership in artificial intelligence. Retrieved from: https://trumpwhitehouse.archives.gov/ai/
58. The White House. (2020, December 3). Executive order on promoting the use of trustworthy artificial intelligence in the federal government. Retrieved from: https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/
59. The White House. (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. Retrieved from: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House
60. Milne, S. (n.d.). Ai Image Generator stable diffusion perpetuates racial and gendered stereotypes, study finds. UW News. https://www.washington.edu/news/2023/11/29/ai-image-generator-stable-diffusion-perpetuates-racial-and-gendered-stereotypes-bias/

61. Hern, A. (2016, September 28). "'Partnership on AI' formed by Google, DeepMind, Facebook, Amazon, IBM and Microsoft | Technology." The Guardian. Retrieved from [https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms]
62. Waters, R. (2016, September 29). "AI is 'Next Big Thing' to worry about." Financial Times. Retrieved from [https://www.ft.com/content/c2c739d0-8661-11e6-8897-2359a58ac7a5]
63. Bindi, T. (2016, September 29). "Amazon, Google, Facebook, IBM, and Microsoft form AI non-profit." ZDNet. Retrieved from [https://www.zdnet.com/article/amazon-google-facebook-ibm-and-microsoft-form-ai-non-profit/]
64. Rubin, B. F., & Cheng, R. (2016, September 29). "The AI Super Friends assemble! (The 3:59, Ep. 115)." CNET. Retrieved from [https://web.archive.org/web/20160930185512/https://www.cnet.com/news/the-ai-super-friends-assemble-the-359-ep-115/]
65. S. J. Ali, A. Christin, A. Smart, and R. Katila. Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23. ACM, June 2023.

66. M. Ashok, R. Madan, A. Joha, and U. Sivarajah. Ethical framework for artificial intelligence and digital technologies. International Journal of Information Management, 62:102433, 2022.

67. J. Ayling and A. Chapman. Putting AI ethics to work: are the tools fit for purpose? AI and Ethics, 2(3):405–429, 2022.

68. J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. Science, 352(6293):1573–1576, 2016.

69. J.-F. Bonnefon, A. Shariff, and I. Rahwan. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE, 107(3):502–504, 2019.

70. A. O. Brightman, N. D. Fila, J. L. Hess, A. J. Kerr, D. Kim, M. C. Loui, and C. B.

    Zoltowski. Applying phenomenography to develop a comprehensive understanding of ethics in engineering practice. In 2018 IEEE Frontiers in Education Conference (FIE), pages 1–5, 2018.

71. T. H. Byers and T. L. Seelig. Empowering future engineers with ethical thinking. The Bridge: Linking Engineering and Society, 50(S):111–114, 2021.

72. N. K. Correa, C. Galvao, J. W. Santos, C. Del Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galv˜ao, E. Terem, and N. de Oliveira. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10):100857,2023.

73. R. Eitel-Porter. Beyond the promise: implementing ethical AI. AI and Ethics, 1:73–80,2021.
74. L. Floridi and J. Cowls. A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, pages 535–545, 2022.
75. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena. Ai4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Machines, 28:689–707, 2018.
76. I. Gabriel. Artificial intelligence, values, and alignment. Minds and Machines, 30:411–437, 09 2020.
77. I. Georgieva, C. Lazo, T. Timan, and A. F. van Veenstra. From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics, 2(4):697–711, 2022.
78. T. Hagendorff. The ethics of AI ethics: An evaluation of guidelines. Minds and machines, 30(1):99–120, 2020.
79. T. Hagendorff. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines,30(1):99–120, Mar. 2020.
80. T. Hagendorff. A virtue-based framework to support putting AI ethics into practice. Philosophy & Technology, 35(3):55, 2022.

81. A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9):389–399, 2019.

82. J.-M. John-Mathews, D. Cardon, and C. Balagu´e. From reality to world. A critical perspective on AI fairness. Journal of Business Ethics, 178(4):945–959, 2022.

83. B. Johnson and J. Smith. Towards ethical data-driven software: Filling the gaps in ethics research & practice. In 2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics), pages 18–25, 2021.

84. N. S. A. Karim, F. A. Ammar, and R. Aziz. Ethical software: Integrating code of ethics into software development life cycle. In 2017 International Conference on Computer and Applications (ICCA), pages 290–298, 2017.

85. A. Martinho. Surveying judges about artificial intelligence: profession, judicial adjudication, and legal principles. AI & SOCIETY, pages 1–16, 2024.

86. A. Martinho, N. Herber, M. Kroesen, and C. Chorus. Ethical issues in focus by the autonomous vehicles industry. Transport Reviews, 41(5):556–577, 2021.

87. A. Martinho, M. Kroesen, and C. Chorus. A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. Artificial intelligence in medicine, 121:102190, 2021.

88. A. Martins Martinho Bessa. Empirical Essays in Artificial Intelligence Ethics. PhD thesis, Delft University of Technology, 2022.

89. B. Mittelstadt. Principles alone cannot guarantee ethical AI—Nature Machine Intelligence, 1(11):501–507, 2019.

90. J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. M ̈okander, and L. Floridi. Ethics as a service: a pragmatic operationalisation of AI ethics. Minds and Machines, 31(2):239 256,2021.

91. J. Morley, L. Floridi, L. Kinsey, and A. Elhalal. From what to how: an initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. Science and engineering ethics, 26(4):2141–2168, 2020.

92. J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi. Operationalising AI ethics: barriers, enablers and next steps. AI & SOCIETY, pages 1–13, 2021.

93. L. Munn. The uselessness of AI ethics. AI and Ethics, 3(3):869–877, 2023.

94. W. Orr and J. L. Davis. Attributions of ethical responsibility by artificial intelligence practitioners. Information, Communication & Society, 23(5):719–735, 2020.

95. E. Prem. From ethical AI frameworks to tools: a review of approaches. AI and Ethics, pages 1–18, 2023.

96. A. Renda. Ethics, algorithms and self-driving cars a csi of the 'trolley problem'. 2018.

97. D. Schiff, J. Biddle, J. Borenstein, and K. Laas. What's next for AI ethics, policy, and governance? a global overview. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pages 153–158, 2020.

98. C. Stix. Actionable principles for artificial intelligence policy: three pathways. Science and engineering ethics, 27(1):15, 2021.

99. A. Theodorou and V. Dignum. Towards ethical and socio-legal governance in AI. Nature Machine Intelligence, 2(1):10–12, 2020.

100. V. Vakkuri, K.-K. Kemell, M. Jantunen, E. Halme, and P. Abrahamsson. Eccola—a method for implementing ethically aligned AI systems. Journal of Systems and Software, 182:111067, 2021.

101. V. Vakkuri, K.-K. Kemell, J. Kultanen, and P. Abrahamsson. The current state of industrial practice in artificial intelligence ethics. IEEE Software, 37(4):50–57, 2020.

102. V. Vakkuri, K.-K. Kemell, J. Tolvanen, M. Jantunen, E. Halme, and P. Abrahamsson. How do software companies deal with artificial intelligence ethics? a gap analysis. In Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, pages 100–109, 2022.

103. C. Wilson and M. Van Der Velden. Sustainable AI: An integrated model to guide public sector decision-making. Technology in Society, 68:101926, 2022.

104. Hendrik R. Schopmans. 2022. From Coded Bias to Existential Threat: Expert Frames and the Epistemic Politics of AI Governance. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 627–640. https://doi.org/10.1145/3514094.3534161

105. Galanos, V. (2018). Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. Technology Analysis & Strategic Management, 31(4), 421–432. https://doi.org/10.1080/09537325.2018.1518521

106. Durkheim E. (1895) 1982. The Rules of Sociological Method. New York: The Free Press.

107. Silvast, A., & Virtanen, M. J. (2023). On Theory–Methods Packages in Science and Technology Studies. Science, Technology, & Human Values, 48(1), 167-189. https://doi.org/10.1177/01622439211040241

108. Gad, C., & Ribes, D. (2014). The conceptual and the empirical in science and technology studies. Science, Technology, & Human Values, 39(2), 183-191.

109. Bracken, L.J. and Oughton, E.A. (2006), 'What do you mean?' The importance of language in developing interdisciplinary research. Transactions of the Institute of British Geographers, 31: 371-382. https://doi.org/10.1111/j.1475-5661.2006.00218.x

110. Cihon, Peter, Jonas Schuett, and Seth D. Baum. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. Information 12, no. 7: 275. https://doi.org/10.3390/info12070275

111. Ali, S. M., Dick, S., Dillon, S., Jones, M. L., Penn, J., & Staley, R. (2023). Histories of artificial intelligence: a genealogy of power. BJHS Themes, 8, 1–18. doi:10.1017/bjt.2023.15

112. Huw Roberts, Emmie Hine, Mariarosaria Taddeo, Luciano Floridi, Global AI governance: barriers and pathways forward, International Affairs, Volume 100, Issue 3, May 2024, Pages 1275–1286, https://doi.org/10.1093/ia/iiae073

113. Le, Quoc & Ranzato, Marc'Aurelio & Monga, Rajat & Devin, Matthieu & Chen, Kai & Corrado, G.s & Dean, Jeff & Ng, Andrew. (2011). Building high-level features using large scale unsupervised learning. Proceedings of ICML. 1.

114. O'Neil, C. (2016). Weapons of math destruction : how big data increases inequality and threatens democracy (First edition.). Crown.

115. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

116. Ryan, M., Antoniou, J., Brooks, L., et al. Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality. Sci Eng Ethics 27, 16 (2021). https://doi.org/10.1007/s11948-021-00293-x

117. Araz Taeihagh, Governance of artificial intelligence, Policy and Society, Volume 40, Issue 2, June 2021, Pages 137–157, https://doi.org/10.1080/14494035.2021.1928377

118. OECD. (2023). National AI Policies & Strategies. In: The OECD Artificial Intelligence Policy Observatory, https://oecd.ai/en/dashboards/overview

119. Menter, D. E. (2013). The Role of Mechanically Induced ERK Phosphorylation in Stem Cell Adipogenesis. https://core.ac.uk/download/188086004.pdf

120. Schuelke-Leech, B.-A., Jordan, S.R. and Barry, B. (2019), Regulating Autonomy: An Assessment of Policy Language for Highly Automated Vehicles. Rev Policy Res, 36: 547-579. https://doi.org/10.1111/ropr.12332

121. Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., & Ng, A. (2012). Building High-level Features Using Large-Scale Unsupervised Learning. In: Proceedings of the 29th International Conference on Machine Learning (ICML). Volume: 28. Pages: 3-10

122. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems (NIPS). Volume: 25. Pages: 1097-1105

123. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

124. Silver, D., Huang, A., Maddison, C. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489 (2016). https://doi.org/10.1038/nature16961

125. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

126. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

127. Bode, I., Huelss, H., Nadibaidze, A. et al. Prospects for the global governance of autonomous weapons: comparing Chinese, Russian, and

US practices. Ethics Inf Technol 25, 5 (2023). https://doi.org/10.1007/s10676-023-09678-x

128. Gordon, JS., Gunkel, D.J. (2024). Artificial Intelligence and the future of work. AI & Soc https://doi.org/10.1007/s00146-024-01960-w

129. Stephen Cave and Seán S. ÓhÉigeartaigh. 2018. An AI Race for Strategic Advantage: Rhetoric and Risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 36–40. https://doi.org/10.1145/3278721.3278780

130. Bostrom, N. (2016). The Control Problem. Excerpts from Superintelligence: Paths, Dangers, Strategies. In Science Fiction and Philosophy, S. Schneider (Ed.). https://doi.org/10.1002/9781118922590.ch23

131. Namdar, B., Pölzler, T. Toby Ord, The Precipice: Existential Risk and the Future of Humanity, Bloomsbury, 2020. Ethic Theory Moral Prac 24, 855–857 (2021). https://doi.org/10.1007/s10677-021-10181-9

132. Shen, Y., Zhang, X. (2024). The impact of artificial intelligence on employment: the role of virtual agglomeration. Humanit Soc Sci Commun 11, 122 . https://doi.org/10.1057/s41599-024-02647-9

133. Schuett, J., Reuel, AK. & Carlier, A. How to design an AI ethics board. AI Ethics (2024). https://doi.org/10.1007/s43681-023-00409-y

134. UNESCO, 2023. Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence. https://doi.org/10.54678/YTSA7796

135. Hirsh, M. (2023). How AI Will Revolutionize Warfare. In: Foreign Policy. Available at: https://foreignpolicy.com/2023/04/11/ai-arms-race-artificial-intelligence-chatgpt-military-technology/

136. Henry A. Kissinger and Graham, Allison. 2023. "The Path to AI Arms Control." Foreign Affairs, October 13. https://www.foreignaffairs.com/united-states/henry-kissinger-path-artificial-intelligence-arms-control.

137. Nick Bostrom. (2012a). Existential Risk Prevention as Global Priority. Global Policy, Vol 4, Issue 1 (2013): 15-31.

138. Russell, Stuart J.; Norvig, Peter (2021). Artificial intelligence: A modern approach (4th ed.). Pearson. pp. 5, 1003. ISBN 9780134610993.

139. Russell, Stuart J. (2020). Human compatible: Artificial intelligence and the problem of control. Penguin Random House. ISBN 9780525558637

140. Nick Bostrom. (2012b). The Superintelligent Will: Motivation And Instrumental Rationality In Advanced Artificial Agents. Minds and Machines, Vol. 22, Iss. 2, May 2012.

141. Shattuck, J., Rama, S., &; Risse, M. (2022). Holding Together The Hijacking of Rights in America and How to Reclaim Them for Everyone. The New Press.

142. Urbina, F., Lentzos, F., Invernizzi, C. et al. Dual use of artificial intelligence-powered drug discovery. Nat Mach Intell 4, 189–191 (2022). https://doi.org/10.1038/s42256-022-00465-9

143. Shermer, Michael (1 March 2017). "Apocalypse AI". Scientific American. 316 (3): 77. Bibcode:2017SciAm.316c..77S. doi:10.1038/scientificamerican0317-77

144. Goldman, S. (2023). "AI experts challenge 'doomer' narrative, including 'extinction risk' claims". VentureBeat. Retrieved 8 June 2024. https://venturebeat.com/ai/ai-experts-challenge-doomer-narrative-including-extinction-risk-claims/

145. Jindal, Siddharth (2023). "OpenAI's Pursuit of AI Alignment is Farfetched". Analytics India Magazine. Retrieved 23 June 2024. https://analyticsindiamag.com/openais-farfetched-pursuit-of-ai-alignment/

146. Roberts, H., Cowls, J., Hine, E., Mazzi, F., Tsamados, A., Taddeo, M., & Floridi, L. (2021). Achieving a ‘Good AI Society’: Comparing the Aims and Progress of the EU and the US. Science and engineering ethics, 27, 1-25.

147. Burnay, M., & Circiumaru, A. (2023). The AI global order: what place for the European Union?. In Contestation and Polarization in Global Governance (pp. 264-281). Edward Elgar Publishing.

148. Iskandarova, M. (2017). From the idea of scale to the idea of agency: An actor-network theory perspective on policy development for renewable energy. Science and Public Policy, 44(4), 476-485.

149. OpenAI. (2022). Introducing ChatGPT. Retrieved 23 June 2024. https://openai.com/index/chatgpt/

150. W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders and K. -R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," in *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247-278, March 2021, doi: 10.1109/JPROC.2021.3060483.

151. Y. -J. Hu and S. -J. Lin, "Deep Reinforcement Learning for Optimizing Finance Portfolio Management," *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates, 2019, pp. 14-20, doi: 10.1109/AICAI.2019.8701368.

152. Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., ... & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and translational science*, *14*(1), 86-93.

153. Yang, T., Yi, X., Lu, S., Johansson, K. H., & Chai, T. (2021). Intelligent manufacturing for the process industry driven by industrial artificial intelligence. *Engineering*, *7*(9), 1224-1230.

154. Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., & Fei-Fei, L. (2013). Detecting avocados to Zucchinis: What have we done, and where are we going? In *Proceedings - 2013 IEEE International Conference on Computer Vision, ICCV 2013* (pp. 2064-2071). Article 6751367 (Proceedings of the IEEE International Conference on Computer Vision). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ICCV.2013.258

155. Lecun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P., & Vapnik, V. (1995). Comparison of learning algorithms for handwritten digit recognition. In F. Fogelman, & P. Gallinari (Eds.), *International Conference on Artificial Neural Networks, Paris* (pp. 53-60). EC2 & Cie.

156. Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

157. Sloane, M. (2024). Controversies, contradiction, and "participation" in AI. Big Data & Society, 11(1). https://doi.org/10.1177/20539517241235862

158. Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). Portrayals and perceptions of AI and why they matter.

159. King, D. K., & Hayes, J. (2017). The effects of power relationships: knowledge, practice and a new form of regulatory capture*. Journal of Risk Research, 21(9), 1104–1116. https://doi.org/10.1080/13669877.2017.1382560

160. Gallagher, R. (2018). *Google CEO Hammered by Members of Congress on China Censorship Plan*. In: The Intercept. Retrieved 23 June 2024. https://theintercept.com/2018/12/11/google-congressional-hearing/

161. BBC News. (2019, July 17). Google's Chinese search engine "terminated." BBC News. Retrieved July 10, 2024, from https://www.bbc.com/news/technology-49015516

162. Copestake, J. (2018, December 18). Google China: Has search firm put Project Dragonfly on hold? BBC News. Retrieved July 10, 2024, from https://www.bbc.com/news/technology-46604085

163. Amnesty International. (2020). Google: Drop Project Dragonfly. Www.amnesty.org.uk. Retrieved July 10, 2024, from https://www.amnesty.org.uk/article-google-drop-dragonfly

164. Pellerin, C. (2017, July 21). Project Maven to Deploy Computer Algorithms to War Zone by Year's End. U.S. Department of Defense. https://www.defense.gov/News/News-Stories/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/

165. Shane, S., & Wakabayashi, D. (2018). "The Business of War": Google Employees Protest Work for the Pentagon. The New York Times. https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

166. European Commission. (2013). Horizon 2020. Research-And-Innovation.ec.europa.eu. Retrieved July 10, 2024, from http://ec.europa.eu/programmes/horizon2020/en/h2020-section/fet-flagships

167. European Parliament. (2016). Procedure File: 2016/2271(INI) | Legislative Observatory | Europa.eu. https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2016/2271(INI)&l=en

168. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) . EUR-Lex - 32022R1925 - EN - EUR-Lex. (2022). Europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925

169. Yasar, Ayse Gizem and Chong, Andrew and Dong, Evan and Gilbert, Thomas and Hladikova, Sarah and Mougan, Carlos and Shen, Xudong and Singh, Shubham and Stoica, Ana-Andreea and Thais, Savannah. (2024). Integration of Generative AI in the Digital Markets Act: Contestability and Fairness from a Cross-Disciplinary Perspective. LSE Legal Studies Working Paper No. 4/2024, Available at SSRN: https://ssrn.com/abstract=4769439

170. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). (2016). Official Journal, L 119, 1-88. ELI: http://data.europa.eu/eli/reg/2016/679/oj[legislation]

171. National Science Foundation. (2011). National Robotics Initiative 3.0: Innovations in Integration of Robotics (NRI-3.0) | NSF. New.nsf.gov.

https://new.nsf.gov/funding/opportunities/national-robotics-initiative-30-innovations/503641/nsf21-559

172. National Institute of Health. (2016). Big Data to Knowledge | NIH Common Fund. Commonfund.nih.gov. https://commonfund.nih.gov/bd2k

173. Kejriwal, M. (2021). Essential Features in a Theory of Context for Enabling Artificial General Intelligence. Applied Sciences, 11(24), 11991.

174. Shead, S. (2017). Apple has confirmed it is joining the Partnership on AI. Business Insider. Retrieved July 10, 2024, from https://businessinsider.com/apple-joins-partnership-on-ai-2017-1

175. Amazon AWS Public Sector Blog. (2023). A framework to mitigate bias and improve outcomes in the new age of AI | AWS Public Sector Blog. Aws.amazon.com. https://aws.amazon.com/blogs/publicsector/framework-mitigate-bias-improve-outcomes-new-age-ai/

176. NSF 19-571: NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon | NSF - National Science Foundation. (2019). New.nsf.gov. Retrieved July 10, 2024, from https://new.nsf.gov/funding/opportunities/nsf-program-fairness-artificial-intelligence/505651/nsf19-571/solicitation

177. Bird, Sarah and Barocas, Solon and Crawford, Kate and Diaz, Fernando and Wallach, Hanna, Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI (October 2, 2016). Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2016, Available at SSRN: https://ssrn.com/abstract=2846909

178. The Future Computed: Artificial Intelligence and its role in society - The Official Microsoft Blog. (2018). The Official Microsoft Blog. https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/

179. OpenAI . (2016). Concrete AI safety problems. Retrieved July 10, 2024, from https://openai.com/index/concrete-ai-safety-problems/

180. Walker, K. (2023). A policy agenda for responsible AI progress: Opportunity, Responsibility, Security. Google. https://blog.google/technology/ai/a-policy-agenda-for-responsible-ai-progress-opportunity-responsibility-security/

181. Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). Natural language processing: python and NLTK. Packt Publishing Ltd.

182. Bogost, I. (2022). ChatGPT Is Dumber Than You Think. The Atlantic.

https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/

183. Roose, K. (2023). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. The New York Times. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

184. Hu, K. (2023). ChatGPT sets record for fastest-growing user base. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

185. Bryson, J. J., & Malikova, H. (2021). Is there an AI Cold War? *Global Perspectives*, 2(1), 24803.

186. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

187. Joachim Roski, Ezekiel J Maier, Kevin Vigilante, Elizabeth A Kane, Michael E Matheny, (2021). Enhancing trust in AI through industry self-governance, Journal of the American Medical Informatics Association, Volume 28, Issue 7, Pages 1582–1590, https://doi.org/10.1093/jamia/ocab065

188. McCarty, Richard. (2009). 'Moral Motivation', *Kant's Theory of Action* (Oxford, 2009; online edn, Oxford Academic, 1 Sept. 2009), https://doi.org/10.1093/acprof:oso/9780199567720.003.0006, accessed 10 July 2024.

189. Kant, I. (1785). Groundwork for the Metaphysics of Morals. Oxford University Press.

190. Altman, S. (2023). Planning for AGI and Beyond. Retrieved July 10, 2024, from https://openai.com/index/planning-for-agi-and-beyond/

191. Sartori, G. (1991). "Comparing and Miscomparing." Journal of Theoretical Politics 3(3): 243-257.

192. De Baere, G., and Gutman, K., (2012). 'Federalism and International Relations in the European Union and the United States: A Comparative Outlook, 'Modern Studies in European Law, Vol. 33, p. 131-166.

193. Dafoe, A. (2015). On Technological Determinism: A Typology, Scope Conditions, and a Mechanism. Science, Technology, & Human Values, 40(6), 1047-1076. https://doi.org/10.1177/0162243915579283

194. Wyatt, S. (2008). Technological Determinism Is Dead; Long Live Technological Determinism. In Hackett E. J., Amsterdamska O., Wajcman J. (Ed.), The Handbook of Science and Technology Studies. Cambridge, MA: MIT Press.

195. Havel, Václav. 1985. The Power of the Powerless: Citizens Against the State in Central-Eastern Europe, edited by J. Keane. Armonk, NY: M. E. Sharpe. ISBN 978-0873327619

196. Jasanoff, Sheila. (2020). "Imagined Worlds: The Politics of Future-Making in the 21st Century." The Politics and Science of Prevision: Governing and Probing the Future. Ed. Andreas Wenger, Ursula Jasper, and Myriam Dunn Cavelty. Routledge.

197. MacKenzie, Donald & Judy Wajcman (eds) ([1985]1999) The Social Shaping of Technology: How the Refrigerator Got Its Hum (Milton Keynes, U.K.: Open University Press); 2nd ed. (1999) (Philadelphia: Open University Press).

198. Marx, L. & Smith, M. R. (eds). (1994). "The Idea of 'Technology' and Postmodern Pessimism,." Does Technology Drive History? The Dilemma of Technological Determinism (Cambridge, MA: MIT Press): 237–58.

199. Bevir, M. (2012). Governance : a very short introduction. Oxford University Press.

200. Fukuyama, F. (2013). "What Is Governance?." CGD Working Paper 314. Washington, DC: Center for Global Development. http://www.cgdev.org/content/publications/detail/1426906

201. The United Nations Development Programme. (1997). Governance for Sustainable Human Development, UNDP Policy Document, New York

202. Bryson, JJ. (2019). The Past Decade and Future of AI's Impact on Society. in Towards a New Enlightenment? A Transcendent Decade. vol. 11, Turner, Madrid.

203. Clement, J. (2024). Google, Amazon, Meta, Apple, and Microsoft (GAMAM) - Statistics & Facts. In: Statista. Retrieved from: https://www.statista.com/topics/4213/google-apple-facebook-amazon-and-microsoft-gafam/#statisticChapter

204. Sørensen, E., & Triantafillou, P. (2016). The politics of self-governance: an introduction. In The politics of self-governance (pp. 1-22). Routledge.

205. Gruetzemacher, R., & Whittlestone, J. (2019). Defining and unpacking transformative AI. arXiv preprint arXiv:1912.00747, 1133.

206. Redmond, Kathy MSc, RN. The US and European Regulatory Systems: A Comparison. Journal of Ambulatory Care Management 27(2):p 105-114, April 2004.

207. Allensworth, R. (2020-2021). Antitrust's High-Tech Exceptionalism. Yale Law Journal Forum, 130, 588-607.

208. Farrell, H. (2003). Constructing the international foundations of e-commerce: The EU-U.S. safe harbor arrangement. International Organization, 57, 277–306.

209. Acemoglu, D., & Lensman, T. (2023). Regulating transformative technologies (No. w31461). National Bureau of Economic Research.

210. Movius, L. B., & Krup, N. (2009). US and EU privacy policy: Comparison of regulatory approaches. International Journal of Communication, 3, 19.

211. Henrikson, A. (August 2016). "Historical Forms of US-European Cooperation: Combination or 'Only' Coordination?," European Foreign Affairs Review, 21, no. 3, pp. 329-356.

212. Takács, T. (2014). "Transatlantic Regulatory Cooperation in Trade: Objectives, Challenges and Instruments for Economic Governance", in Fahey, Elaine & Deirdre Curtin (eds), A Transatlantic Community of Law: Legal Perspectives on the Relationship between the EU and US Legal Orders, Cambridge, Cambridge University Press, 2014, pp. 158-185.

213. Riccardo Crescenzi, Andrés Rodríguez-Pose, Michael Storper, The territorial dynamics of innovation: a Europe–United States comparative analysis, Journal of Economic Geography, Volume 7, Issue 6, November 2007, Pages 673–709, https://doi.org/10.1093/jeg/lbm030

214. Matthijs, M, and Parsons, C, “Single-Market Power: How Europe Surpassed America in the Quest for Economic Integration,” FOREIGN AFFAIRS, May-June 2022, pp. 165-176.

215. Komaitis, K, and J Sherman, “US and EU tech strategies aren’t as aligned as you think,” Brookings Institution, available at: https://www.brookings.edu/techstream/us-and-eu-tech-strategy-arent-as-aligned-as-you-think/.

216. Engler, A. (2023). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings. https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

217. Summers, C. W. (2001). Individualism, Collectivism and Autonomy in American Labor Law. Emp. Rts. & Emp. Pol'y J., 5, 453.

218. Keynes J. M. (1936). The General Theory of Employment, Interest and Money. Palgrave Macmillan.

219. Rosenthal, L. H., & Joseph, G. P. (2017). Foundations of US Federalism. Judicature, 101, 39.

220. Greif, A., Nye, J. V., & Kiesling, L. (2020). Institutions, innovation, and industrialization: Essays in economic history and development.

221. Prové, C. (2018). The politics of urban agriculture : An international exploration of governance, food systems, and environmental justice. https://core.ac.uk/download/158345144.pdf

222. Craig, R. K., Garmestani, A. S., Allen, C. R., Arnold, C. A. T., Birgé, H., DeCaro, D. A., ... & Schlager, E. (2017). Balancing stability and flexibility in adaptive governance: an analysis of tools available in US environmental law. Ecology and society: a journal of integrative science for resilience and sustainability, 22(2), 1.

223. Atkinson, Robert D. and Andes, Scott M., The Atlantic Century: Benchmarking EU & U.S. Innovation and Competitiveness (February 25, 2009). Available at SSRN: https://ssrn.com/abstract=1435523

224. Borrás, S. (2003). Innovation Policies in Europe and the US: The New Agenda (P.S. Biegelbauer, Ed.) (1st ed.). Routledge. https://doi.org/10.4324/9781315195902

225. Orbie, J., & Babarinde, O. (2013). The social dimension of globalization and EU development policy: promoting core labour standards and corporate social responsibility. In Policy Coherence and EU Development Policy (pp. 133-151). Routledge.

226. Carpenter, D., & Moss, D. A. (Eds.). (2013). Preventing regulatory capture: Special interest influence and how to limit it. Cambridge University Press.

227. Brattberg, E., & Rhinard, M. (2011). Multilevel governance and complex threats: The case of pandemic preparedness in the European Union and the United States. Global Health Governance, 5(1), 1-21.

228. Hix, S., & Høyland, B. (2022). The political system of the European Union. Bloomsbury Publishing.

229. Dennett, D. C. (1989). The intentional stance. Cambridge, MA: MIT Press.

230. Wendt, A. E. (1987). The agent-structure problem in international relations theory. International Organization, 41(3), 335–370. doi:10.1017/S002081830002751X

231. Fabbrini, S. (2010). Compound democracies: Why the United States and Europe are becoming similar. OUP Oxford.

232. Slaughter, S., & Cantwell, B. (2012). Transatlantic moves to the market: The United States and the European Union. Higher education, 63, 583-606.

233. Doctorow, C. (2023). How Big Tech Got So Damn Big. WIRED; WIRED.

234. Jones, Meg, Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw (June 1, 2017). Journal of Law, Technology & Policy (Fall 2018), Available at SSRN: https://ssrn.com/abstract=2981855

235. Coll-Mayor, D., Paget, M., & Lightner, E. (2007). Future intelligent power grids: Analysis of the vision in the European Union and the United States. Energy Policy, 35(4), 2453-2465.

236. Bendiek, A., & Stürzer, I. (2022). Advancing European internal and external digital sovereignty: The Brussels effect and the EU-US Trade and Technology Council.

237. Moser, C. (2023). Some Reflections on Anticipatory Governance in EU Defence. Max Planck Institute for Comparative Public Law & International Law (MPIL) Research Paper, (2023-14), 165-190.

238. Bahçecik, Ş. O. (2019). Civil society responds to the AWS: Growing activist networks and shifting frames. Global Policy, 10(3), 365-369.

239. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.

240. Rawls, J. (1985). Justice as fairness. Irvington.

Appendix A: AI Policy Landscape in the United States Since 2010[16]

Policy Initiative	Start date	Responsible organization(s)
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	2023	The White House
NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE	2023	National Science Foundation (NSF)|Office of Science and Technology Policy (OSTP)
AI CLAUSE IN EXECUTIVE ORDER ON ADVANCING RACIAL EQUITY AND SUPPORT FOR UNDERSERVED COMMUNITIES THROUGH THE FEDERAL GOVERNMENT	2023	The White House
Voluntary Commitments from leading AI companies	2023	The White House
NIST AI Risk Management Framework	2023	National Institute of Standards and Technology (NIST)
NIST AI Public Working Groups	2023	U.S. Department of Commerce

[16] source: oecd.ai, thewhitehouse.gov

PCAST Generative AI Working Group	2023	President's Council of Advisors on Science and Technology
BLUEPRINT FOR AN AI BILL OF RIGHTS	2022	The White House |Office of Science and Technology Policy (OSTP)
NATIONAL DEFENSE AUTHORIZATION ACT FOR FISCAL YEAR 2022	2022	Department of Defense (DOD)
CHIPS FOR AMERICA ACT	2022	Department of Commerce (DOC)
AI RESEARCHERS' PORTAL	2022	National AI Initiative Office
CHIEF DIGITAL AND AI OFFICE	2022	Department of Defense (DOD)
HUMAN INTERPRETABLE ATTRIBUTION OF TEXT USING UNDERLYING STRUCTURE PROGRAM	2022	Intelligence Advanced Research Projects Activity (IARPA; Intelligence Advanced Research Projects Activity)
Automated Employment Decision Tool Law	2022	New York City Council
New York, Electronic Monitoring Bill	2022	New York State Legislature
NITRD-NAIIO SUPPLEMENT TO THE PRESIDENT'S FY2022 BUDGET	2021	National Science and Technology Council (NSTC)|Subcommittee on Networking & Information Technology Research & Development

		(NITRD)|Office of Science and Technology Policy (OSTP)
NATIONAL AI INITIATIVE OFFICE	2021	The White House |Office of Science and Technology Policy (OSTP)
NATIONAL AI RESEARCH RESOURCE TASK FORCE	2021	The White House |National Science Foundation (NSF)
FEDERAL TRADE COMMISSION CONSUMER PROTECTION AND COMPETITION INVESTIGATIONS (BIAS IN ALGORITHMS AND BIOMETRICS)	2021	Federal Trade Commission (FTC; Federal Trade Commission)
NATIONAL DEFENSE AUTHORIZATION ACT FOR FISCAL YEAR 2021	2021	The White House |Department of Defense (DOD)
AI/ML-BASED SOFTWARE AS A MEDICAL DEVICE ACTION PLAN	2021	Food and Drug Administration (FDA; Food and Drug Administration)
QUAD PRINCIPLES ON TECHNOLOGY DESIGN, DEVELOPMENT, GOVERNANCE, AND USE	2021	The White House
AI TRAINING FOR THE ACQUISITION WORKFORCE ACT (BILL S-2551)	2021	The White House |Office of Management and Budget (OMB)|General Services Administration

REQUEST FOR INFORMATION AND COMMENT ON FINANCIAL INSTITUTIONS USE OF AI, INCLUDING ML	2021	Department of the Treasury |Federal Reserve |Federal Deposit Insurance Corporation |Consumer Financial Protection Bureau (CFPB)|National Credit Union Administration
NSF PROGRAM ON FAIRNESS IN AI	2021	National Science Foundation (NSF)
U.S.-EU TRADE AND TECHNOLOGY COUNCIL	2021	European Commission Executive Vice-President Margrethe Vestager, European Commission Executive Vice-President Valdis Dombrovskis, US Secretary of State Antony Blinken, US Secretary of Commerce Gina Raimondo, and US Trade Representative Katherine Tai
Artificial Intelligence and Algorithmic Fairness Initiative	2021	Equal Employment Opportunity Commission (EEOC)
DECLARATION OF U.S.-UK CO-OPERATION IN AI R&D	2020	The White House
NATIONAL AI RESEARCH INSTITUTES	2020	National Science Foundation (NSF)

U.S. AI COVID-19 RESPONSE	2020	Office of Science and Technology Policy (OSTP)|Department of Energy (DOE)
NIST PRINCIPLES FOR EXPLAINABLE AI	2020	National Institute of Standards and Technology (NIST)
U.S. PATENT AND TRADEMARK OFFICE REPORT ON PUBLIC VIEWS ON AI AND INTELLECTUAL PROPERTY POLICY	2020	United States Patent and Trademark Office (USPTO)
NATIONAL STRATEGY FOR CRITICAL AND EMERGING TECHNOLOGIES	2020	The White House
GOVERNMENT BY ALGORITHM: AI IN FEDERAL ADMINISTRATIVE AGENCIES	2020	Administrative Conference of the United States (ACUS)
EXECUTIVE ORDER ON PROMOTING THE USE OF TRUSTWORTHY AI IN FEDERAL GOVERNMENT	2020	Office of Science and Technology Policy (OSTP)|The White House
STATE DEPARTMENT GUIDANCE ON PRODUCTS OR SERVICES WITH SURVEILLANCE CAPABILITIES	2020	Department of State (DOS)
NATIONAL AI INITIATIVE ACT OF 2020	2020	Office of Science and Technology Policy (OSTP)|The White House

NATIONAL AI ADVISORY COMMITTEE	2020	Office of Science and Technology Policy (OSTP)|The White House
NATIONAL AI ADVISORY COMMITTEE SUBCOMMITTEE ON LAW ENFORCEMENT	2020	Office of Science and Technology Policy (OSTP)|The White House
Illinois Artificial Intelligence Video Interview Act	2020	Illinois State Legislature
Maryland Facial Recognition Law	2020	Maryland State Legislature
EXECUTIVE ORDER ON MAINTAINING AMERICAN LEADERSHIP IN AI	2019	Office of Science and Technology Policy (OSTP)
NATIONAL AI R&D STRATEGIC PLAN: 2019 UPDATE	2019	Select Committee on AI (SCAI)|National Science and Technology Council (NSTC)|Subcommittee on Networking & Information Technology Research & Development (NITRD)
FEDERAL DATA STRATEGY	2019	Federal Geospatial Data Committee |Presidents Management Council |General Services Administration |National Center for Education Statistics, Department of Education and Training (DET)|Federal Statistical

		Research Data Center Program Management Office |U.S. Census Bureau |Department of Commerce (DOC)|Federal Committee on Statistical Methodology |Interagency Council on Statistical Policy |Department of Education (ED)|Office of Management and Budget (OMB)|Interagency Council on Statistical Policy
PLAN FOR FEDERAL ENGAGEMENT IN DEVELOPING TECHNICAL STANDARDS AND RELATED TOOLS	2019	Department of Commerce (DOC)|Interagency Council on Statistical Policy |National Institute of Standards and Technology (NIST)
AMERICAN WORKFORCE POLICY ADVISORY BOARD	2019	Department of Commerce (DOC)
R&D WORKFORCE TRAINING	2019	National Science Foundation (NSF)|National Institute of Standards and Technology (NIST)|Department of Defense (DOD)|Department of

		State (DOS)|Department of Agriculture (USDA)
REQUEST FOR INFORMATION FOR IDENTIFYING PRIORITY ACCESS OR QUALITY IMPROVEMENTS FOR FEDERAL DATA AND MODELS FOR AI R&D AND TESTING	2019	Office of Management and Budget (OMB)
STRATEGY FOR AUGMENTING INTELLIGENCE USING MACHINES	2019	Office of the Director of National Intelligence (ODNI)
PROPOSED REGULATORY FRAMEWORK FOR MODIFICATIONS TO AI/ML-BASED SOFTWARE AS A MEDICAL DEVICE	2019	Food and Drug Administration (FDA)
DEPARTMENT OF ENERGY AI AND TECHNOLOGY OFFICE	2019	Department of Energy (DOE)
WHITE HOUSE SUMMIT ON AI IN GOVERNMENT	2019	Office of Science and Technology Policy (OSTP)
MEMORANDUM TO HEADS OF AGENCIES ON REGULATORY AND NON-REGULATORY APPROACHES TO AI	2019	Office of Management and Budget (OMB)|Office of Science and Technology Policy (OSTP)
LOCAL, STATE, AND FEDERAL REGULATIONS ON FACIAL RECOGNITION TECHNOLOGIES	2019	Homeland Security and Governmental Affairs |U.S. State Governments
DEFENSE INNOVATION BOARD AI PRINCIPLES	2019	Department of Defense (DOD)

ADDITION OF SOFTWARE SPECIALLY DESIGNED TO AUTOMATE THE ANALYSIS OF GEOSPATIAL IMAGERY TO THE EXPORT CONTROL CLASSIFICATION NUMBER 0Y521 SERIES	2019	Department of Commerce (DOC)|Department of Defense (DOD)|Department of State (DOS)
NITRD SUPPLEMENT TO THE PRESIDENT'S FY2020 BUDGET	2019	National Science and Technology Council (NSTC)|Subcommittee on Networking & Information Technology Research & Development
PROTECTING THE U.S. ADVANTAGE IN AI AND RELATED CRITICAL TECHNOLOGIES	2018	Department of Defense (DOD)|National Security Council (NSC)
SELECT COMMITTEE ON AI	2018	Office of Science and Technology Policy (OSTP)|The White House
NSF FUNDING OPPORTUNITIES WITH SPECIAL EMPHASIS ON AI	2018	National Science Foundation (NSF)
FEDERAL 5-YEAR STEM EDUCATION STRATEGIC PLAN	2018	National Science and Technology Council (NSTC)|Committee on STEM Education (CoSTEM)|Department of Education (ED)
DEPARTMENT OF DEFENSE AI STRATEGY	2018	Department of Defense (DOD)|U.S. Air Force (USAF)

JOINT AI CENTER	2018	Department of Defense (DOD)
NATIONAL SECURITY COMMISSION ON AI	2018	National Security Council on AI (NSCAI)
WHITE HOUSE SUMMIT ON AI FOR AMERICAN INDUSTRY	2018	Office of Science and Technology Policy (OSTP)
POLICIES THAT REGULATE FINTECH INNOVATION	2018	Consumer Financial Protection Bureau (CFPB)
"AI NEXT" CAMPAIGN	2018	Defense Advanced Research Projects Agency (DARPA)
NEXT-GENERATION NON-SURGICAL NEUROTECHNOLOGY PROGRAMME	2018	DARPA
EXPLAINABLE AI PROGRAM	2018	Defense Advanced Research Projects Agency (DARPA)
ML AND AI SUBCOMMITTEE	2018	Office of Science and Technology Policy (OSTP)|National Science Foundation (NSF)|Department of Commerce (DOC)|Department of Energy (DOE)
AI R&D INTERAGENCY WORKING GROUP	2018	Subcommittee on Networking & Information Technology Research & Development (NITRD)

UNMANNED AIRCRAFT SYSTEMS INTEGRATION PILOT PROGRAM	2017	Federal Aviation Administration (FAA)
AUTOMATED VEHICLES 3.0: PREPARING FOR THE FUTURE OF TRANSPORTATION	2017	Department of Transportation (DOT)
FEDERAL AUTOMATED VEHICLES POLICY	2016	Department of Transportation (DOT)
BIG DATA TO KNOWLEDGE	2016	National Institutes of Health (NIH)
NATIONAL AI R&D STRATEGIC PLAN	2016	National Science and Technology Council (NSTC)|Office of Science and Technology Policy (OSTP)|Subcommittee on Networking & Information Technology Research & Development (NITRD)
ACCELERATING MEDICINES PARTNERSHIP	2014	National Institutes of Health (NIH)|Food and Drug Administration (FDA)
CENTERS OF EXCELLENCE IN REGULATORY SCIENCE AND INNOVATION	2013	Food and Drug Administration (FDA)
MATERIALS GENOME INITIATIVE	2011	National Science and Technology Council (NSTC)
NATIONAL ROBOTICS INITIATIVE	2011	National Science Foundation (NSF)|Department of Agriculture

		(USDA)|Department of Energy (DOE)|National Aeronautics and Space Administration (NASA)|Department of Defense (DOD)

Appendix B: AI Policy Landscape in the European Union Since 2010[17]

Policy Initiative	Start date	Responsible organization(s)
European Centre for Algorithmic Transparency	2023	European Commission (EC)|Joint Research Center (JRC)
Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme	2023	European Commission
EUROPEAN DIGITAL INFRASTRUCTURE CONSORTIUM (EDIC)	2023	European Commission (EC)
ASSESSMENT OF CURRENT INITIATIVES OF THE EUROPEAN COMMISSION ON BETTER REGULATION	2022	European Parliament Committee on Legal Affairs (JURI)
EUROPEAN LIGHTHOUSE ON SECURE AND SAFE AI	2022	European Laboratory for Learning and Intelligent Systems (ELLIS)
AI, DATA, AND ROBOTICS PARTNERSHIP IN HORIZON EUROPE	2021	European Commission (EC)
DESTINATION EARTH INITIATIVE	2021	European Commission (EC)

[17] source: oecd.ai, https://commission.europa.eu/

EUROPEAN PUBLIC PROCUREMENT DATA STRATEGY	2021	DG GROW
AI ACT	2021	European Commission (EC)
HORIZON EUROPE	2020	European Commission (EC)
FINANCING OF ARTIFICIAL INTELLIGENCE AND BLOCKCHAIN TECHNOLOGIES	2020	European Commission (EC)|European Investment Bank (EIB)
WHITE PAPER ON ARTIFICIAL INTELLIGENCE	2020	European Commission (EC)
DIGITAL SERVICES ACT PACKAGE	2020	European Commission (EC)
DIGITAL EDUCATION ACTION PLAN	2020	European Commission (EC)
RESOLUTION ON AUTOMATED DECISION-MAKING PROCESSES	2020	European Parliament (EP)|European Parliament Committee on Legal Affairs (JURI)
SPECIAL COMMITTEE ON ARTIFICIAL INTELLIGENCE IN THE DIGITAL AGE	2020	European Parliament (EP)
JURI REPORTS ON "MAKING AI EUROPEAN"	2020	European Parliament (EP)
REPORT ON SAFETY AND LIABILITY IMPLICATIONS OF ARTIFICIAL INTELLIGENCE, THE INTERNET OF THINGS AND ROBOTICS	2020	European Commission (EC)

EUROPEAN STRATEGY FOR DATA	2020	European Commission (EC)
THE ROBUSTNESS AND EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE	2020	European Commission (EC)
COUNCIL CONCLUSIONS ON REGULATORY SANDBOXES	2020	Council of the European Union
SHAPING DIGITAL EDUCATION POLICY	2020	European Parliament (EP; European Parliament)
AI IN EDUCATION, CULTURE AND THE AUDIOVISUAL SECTOR	2020	European Parliament (EP)
CLOSING THE DIGITAL GENDER GAP: WOMENS PARTICIPATION IN THE DIGITAL ECONOMY	2020	European Parliament (EP)
CIVIL LIABILITY REGIME FOR ARTIFICIAL INTELLIGENCE	2020	European Parliament (EP)
FRAMEWORK OF ETHICAL ASPECTS OF ARTIFICIAL INTELLIGENCE, ROBOTICS AND RELATED TECHNOLOGIES	2020	European Parliament (EP)
RESOLUTION ON INTELLECTUAL PROPERTY RIGHTS FOR THE DEVELOPMENT OF AI TECHNOLOGIES	2020	European Parliament (EP)
RESOLUTION ON AI: QUESTIONS OF INTERPRETATION AND APPLICATION OF INTERNATIONAL LAW	2020	European Parliament (EP)

DATA GOVERNANCE ACT	2020	European Commission (EC)
COUNCIL CONCLUSIONS ON THE COORDINATED PLAN ON THE DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE MADE IN EUROPE	2019	Council of the European Union
COMMON EUROPEAN DATA SPACE	2019	European Commission (EC)
POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE	2019	European Commission (EC)
OPEN DATA DIRECTIVE	2019	European Commission (EC)
EUROPEAN NETWORK FOR AI EXCELLENCE CENTRES	2019	European Commission (EC)
SCIENTIFIC ADVICE TO EUROPEAN POLICY IN A COMPLEX WORLD	2019	SAM
EUROPEAN HIGH PERFORMANCE COMPUTING JOINT UNDERTAKING (EuroHPC JU)	2019	European Union Joint Undertaking
AI4EU CONSORTIUM	2019	AI4EU
DIGITAL EUROPE PROGRAMME	2019	European Commission (EC)
AI WATCH	2019	European Commission (EC)
RESOLUTION ON A COMPREHENSIVE EUROPEAN INDUSTRIAL POLICY ON	2019	European Parliament (EP)|European Parliament

ARTIFICIAL INTELLIGENCE AND ROBOTICS		Committee on Legal Affairs (JURI)
RESOLUTION ON AUTONOMOUS DRIVING IN EUROPEAN TRANSPORT	2019	European Parliament (EP)
STOA CENTRE FOR AI	2019	European Parliament (EP)
ADVANCED TECHNOLOGIES FOR INDUSTRY	2019	European Commission (EC)
EU COMMUNICATION ON ARTIFICIAL INTELLIGENCE	2018	European Commission (EC)
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE	2018	European Commission (EC)
THE EUROPEAN AI ALLIANCE	2018	European Commission (EC)
COORDINATED PLAN ON ARTIFICIAL INTELLIGENCE	2018	European Commission (EC)
EU STRATEGY FOR ARTIFICIAL INTELLIGENCE	2018	European Commission (EC)
DECLARATION OF COOPERATION ON AI	2018	European Commission (EC)
ETHICS GUIDELINES ON ARTIFICIAL INTELLIGENCE	2018	European Commission (EC)
GENERAL DATA PROTECTION REGULATION	2018	European Parliament (EP)|Council of the European Union
INNOVATIVE PUBLIC SERVICES	2018	European Commission (EC)
RESOLUTION ON CIVIL LAW RULES ON ROBOTICS AND AI	2017	European Parliament (EP)|European Parliament

		Committee on Legal Affairs (JURI)
RESOLUTION ON DIGITISING EUROPEAN INDUSTRY	2017	European Parliament (EP)|European Parliament Committee on Legal Affairs (JURI)
DIGITAL INNOVATION HUBS	2016	European Commission (EC)
EU CLOUD INITIATIVE	2016	European Commission (EC)
ROBOTICS PUBLIC-PRIVATE PARTNERSHIP IN HORIZON 2020	2014	European Commission (EC)
PUBLIC-PRIVATE PARTNERSHIP ON DATA	2014	European Commission (EC)
BIG DATA VALUE PPP	2014	European Commission (EC)
HORIZON 2020	2014	European Commission (EC)
FET FLAGSHIP- HUMAN BRAIN PROJECT	2013	European Commission (EC)

Appendix C: Non-State Actors' AI Policy Landscape Since 2010

Policy Initiative	Start Date	Responsible Organization(s)	Retrieved from
Advancing positive outcomes for people and society	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/about/
Announcing PAI's Policy Steering Committee	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/pai-announces-policy-steering-committee/
Partnership onAI response to the Request for Information (RFI)	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon,	www.partnershiponai.org

Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence		Microsoft, IBM, Apple)	
The Time Is Now to Act Together on AI Governance - Partnership on AI	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/pais-policy-forum-the-time-is-now-to-act-together-on-ai-governance/
Responsible AI Recommendations	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/responsible-ai-recommendations-top-four-pai-resources-of-2023/

AI Governance Requires Global Action	2023	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/ai-governance-requires-global-action/
PAI Developing Ethical Guidelines for Synthetic Media	2022	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/pai-developing-ethical-guidelines-for-synthetic-media/
PAI Annual Report	2022	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	www.partnershiponai.org
Supporting Corporate Responsibilities with	2022	Partnership on AI (Facebook, Google and its DeepMind subsidiary,	https://partnershiponai.org/supporting-corporate-responsibilities-with-emerging-ai-technologies-insights-from-the-pai-board/

Emerging AI Technologies		Amazon, Microsoft, IBM, Apple)	
The Partnership on AI Response to the National Institute of Standards and Technology R equest for Information - AI Risk Management Framework	2021	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://www.nist.gov/document/nist-ai-rfi-partnershiponai001pdf
PAI Annual Report	2021	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	www.partnershiponai.org

The Partnership on AI Response to the White House Office of Science and Technology Policy and National Science Foundation Request for Information: National Artificial Intelligence Research Resource	2021	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/pai-submits-response-to-ostp-nsf-request-for-information-on-national-ai-research-resource/
From Affirmative Action to Affirmative Algorithms: The Legal Challenges Threatening Progress on	2020	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/affirmativealgorithms/

Algorithmic Fairness			
From Insight to Impact: Reflections From the 2020 All Partners Meeting	2020	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/from-insight-to-impact-reflections-from-the-2020-all-partners-meeting/
LETTER FROM THE FOUNDING EXECUTIVE DIRECTOR	2020	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	www.partnershiponai.org
THE RESPONSIBLE AI ART FIELD GUIDE	2020	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://www.partnershiponai.org/
Co-Creating the Future of Responsible	2019	Partnership on AI (Facebook, Google and its	https://partnershiponai.org/reflections-from-apm/

AI: Reflections from the All Partners Meeting		DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	
Working Group Charters: Guiding our exploration of AI's hard questions	2018	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions/
Partnership On AI Strengthens Its Network of Partners and Announces First Initiatives	2017	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon, Microsoft, IBM, Apple)	https://partnershiponai.org/partnership-on-ai-strengthens-its-network-of-partners-and-announces-first-initiatives/
Announcing the Partnership on AI	2016	Partnership on AI (Facebook, Google and its DeepMind subsidiary, Amazon,	https://www.partnershiponai.org/the-partnership-on-ai-launches-multistakeholder-initiative-to-enhance-machine-learning-

		Microsoft, and IBM,)	
Democratic inputs to AI	2023	OpenAI	https://openai.com/blog/democratic-inputs-to-ai
Governance of superintelligence	2023	OpenAI	https://openai.com/blog/governance-of-superintelligence
How should AI systems behave, and who should decide?	2023	OpenAI	https://openai.com/blog/how-should-ai-systems-behave
Our approach to AI safety	2023	OpenAI	https://openai.com/blog/our-approach-to-ai-safety
Planning for AGI and beyond	2023	OpenAI	https://openai.com/blog/planning-for-agi-and-beyond
Safety standards	2023	OpenAI	https://openai.com/safety-standards
Weak-to-strong generalization	2023	OpenAI	https://openai.com/research/weak-to-strong-generalization
Best practices for deploying	2022	OpenAI	https://openai.com/blog/best-practices-for-deploying-language-models

language models			
Economic impacts research at OpenAI	2022	OpenAI	https://openai.com/blog/economic-impacts
Our approach to alignment research	2022	OpenAI	https://openai.com/blog/our-approach-to-alignment-research
The Role of Cooperation in Responsible AI Development	2019	OpenAI	arXiv:1907.04534v1
AI Safety Needs Social Scientists	2019	OpenAI	https://distill.pub/2019/safety-needs-social-scientists/
OpenAI Charter	2018	OpenAI	https://openai.com/charter
Introducing OpenAI	2015	OpenAI	https://openai.com/blog/introducing-openai
Ethical AI in Insurance White Paper	2017	Microsoft, Johns Hopkins University, Swiss Re	https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWFKCm

Frontier Model Forum	2023	Microsoft, Anthropic, Google, and OpenAI	https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/
Machine Learning for fair decisions	2018	Microsoft Research	https://www.microsoft.com/en-us/research/blog/machine-learning-for-fair-decisions/?culture=en-us&country=us 1/7
Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI	2016	Microsoft Research	http://ssrn.com/abstract=2846909
Ethical Considerations for the Use of AI	2018	Microsoft Corporation	https://www.microsoft.com/cms/api/am/binary/RE4pKH5#:~:text=4-,Microsoft%20AI%20guiding%20principles,inclusiveness%2C%20transparency%2C%20and%20accountability.
How do we best govern AI?	2023	Microsoft	https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai/?culture=en-us&country=us
Voluntary Commitment	2023	Microsoft	https://blogs.microsoft.com/wp-content/uploads/prod/sites

s by Microsoft to Advance Responsible AI Innovation			/5/2023/07/Microsoft-Voluntary-Commitments-July-21-2023.pdf
Our commitments to advance safe, secure, and trustworthy AI	2023	Microsoft	https://blogs.microsoft.com/on-the-issues/2023/07/21/commitment-safe-secure-ai/
Microsoft Responsible AI Standard	2022	Microsoft	https://aka.ms/ResponsibleAIQuestions
Age of Intelligence Democratizing AI to empower individuals, organizations and society towards fulfilling the promise of holistic growth	2022	Microsoft	https://www.microsoft.com/en-us/ai/ai-for-hu-manitarian-action

Artificial Intelligence in the Public Sector European Outlook for 2020 and Beyond	2020	Microsoft	https://www.microsoft.com/en-us/ai/responsible-ai-resources
Putting principles into practice: How we approach responsible AI at Microsoft	2020	Microsoft	https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/
Companies should focus on AI ethics	2018	Microsoft	https://news.microsoft.com/wp-content/uploads/prod/sites/68/2018/12/Hugh.jpg
The Future Computed	2018	Microsoft	https://news.microsoft.com/cloudforgood/_media/downloads/the-future-computed-english.pdf
Microsoft AI: Empowering transformation	2018	Microsoft	https://blogs.microsoft.com/ai/microsoft-ai-empowering-transformation/

Technology, ethics and the law: Grappling with our AI-powered future	2018	Microsoft	https://news.microsoft.com/apac/features/technology-ethics-and-the-law-grappling-with-our-ai-powered-future/
Responsible AI Dashboard Components - Microsoft Responsible AI	2018	Microsoft	https://responsibleaitoolbox.ai/introducing-responsible-ai-dashboard/#dashboard-components
Why we launched DeepMind Ethics & Society	2017	Google DeepMind	https://deepmind.google/discover/blog/why-we-launched-deepmind-ethics-society/#:~:text=To%20guarantee%20the%20rigor%2C%20transparency,other%20academics%20and%20civil%20society.
Artificial Intelligence, Values and Alignment	2017	Google DeepMind	https://deepmind.google/discover/blog/artificial-intelligence-values-and-alignment/
AI Principles Progress Update 2023	2023	Google	https://ai.google/static/documents/ai-principles-2023-progress-update.pdf

3 emerging practices for responsible generative AI	2023	Google	https://blog.google/technology/ai/google-responsible-generative-ai-best-practices/
2022 AI Principles Progress Update	2022	Google	https://ai.google/responsibility/principles/
2021 AI Principles Progress Update	2021	Google	https://ai.google/responsibility/principles/
AI Principles 2020 Progress update	2020	Google	https://ai.google/responsibility/principles/
AI Principles 1-Year Progress Update	2019	Google	https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/
An external advisory council to help advance the responsible development of AI	2019	Google	https://blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/

Recommendations for Regulating AI	2019	Google	https://ai.google/static/documents/recommendations-for-regulating-ai.pdf
AI at Google: our principles	2018	Google	https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/
Perspectives on Issues in AI Governance	2018	Google	https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf
Google I/O Transforming democracy and disasters with APIs	2014	Google	g.co/io14videos
DeepMind Mission, Values, Culture & Jobs	2010 - 2015	DeepMind	https://www.tealhq.com/company/deepmind
Ethics and Compliance	2021 - 2023	Apple	https://www.apple.com/compliance/
Ethics and Compliance - Policies	2021 - 2023	Apple	https://www.apple.com/compliance/policies/

NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon	2019	Amazon, National Science Foundation	https://new.nsf.gov/funding/opportunities/nsf-program-fairness-artificial-intelligence/505651/nsf19-571/solicitation
A framework to mitigate bias and improve outcomes in the new age of AI	2023	Amazon	https://aws.amazon.com/blogs/publicsector/framework-mitigate-bias-improve-outcomes-new-age-ai/
Amazon announces commitment to responsible AI	2023	Amazon	https://www.aboutamazon.com/news/company-news/amazon-responsible-ai
AWS Reaffirms its Commitment to Responsible Generative AI	2023	Amazon	https://aws.amazon.com/blogs/machine-learning/aws-reaffirms-its-commitment-to-responsible-generative-ai/

AWS Responsible AI Policy	2023	Amazon	https://aws.amazon.com/machine-learning/responsible-ai/policy/
Responsible AI – Building AI responsibly at AWS	2023	Amazon	https://aws.amazon.com/machine-learning/responsible-ai/
Responsible AI in the generative era	2023	Amazon	https://www.amazon.science/blog/responsible-ai-in-the-generative-era
Fairness, accountability, transparency, ethics	2022	Amazon	https://www.amazon.science/tag/fairness-accountability-transparency-ethics-fate
Introducing AWS AI Service Cards: A new resource to enhance transparency and advance responsible AI	2022	Amazon	https://aws.amazon.com/blogs/machine-learning/introducing-aws-ai-service-cards-a-new-resource-to-enhance-transparency-and-advance-responsible-ai/

Protecting Consumers and Promoting Innovation – AI Regulation and Building Trust in Responsible AI	2022	Amazon	https://aws.amazon.com/blogs/machine-learning/protecting-consumers-and-promoting-innovation-ai-regulation-and-building-trust-in-responsible-ai/
reMARS revisited: Frontiers of fair and accessible AI	2022	Amazon	https://www.amazon.science/latest-news/remars-revisited-frontiers-of-fair-and-accessible-ai
Amazon AI Fairness and Explainability Whitepaper	2022	Amazon	https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf
Some Thoughts on Facial Recognition Legislation	2019	Amazon	https://aws.amazon.com/blogs/machine-learning/some-thoughts-on-facial-recognition-legislation/?linkCode=w50&tag=w050b-20&imprToken=cEv5iz3t1KoAHE3

Appendix D: AI Moral Landscape Dictionary

Theme	Related Keywords
Algorithmic_Decision-Making	ai safety, algorithm, decision boundaries, interpretability, machine learning, model accountability, model bias, model deployment, model drift, model evaluation, model explainability, model fairness, model fairness metrics, model interpretability, model interpretability techniques, model monitoring, model robustness, model selection, model transparency, model validation, predictive models
Data_Governance	anonymization, consent, data, data access controls, data audits, data breach, data ethics, data ethics boards, data governance frameworks, data minimization, data ownership, data portability, data privacy impact assessments, data provenance, data protection, data quality, data retention, data sharing, data sovereignty, data stewardship, data localization, privacy
Defense_and_State_Security	adversarial attacks, ai in defense, arms control, autonomous drones, counterterrorism, critical infrastructure protection, cybersecurity, cyberwarfare, defense, defense policy, dual-use technology, geopolitical stability, intelligence sharing, military applications, military readiness, national resilience, national security, nuclear

	deterrence, secure ai deployment, secure development, strategic alliance, surveillance
Education	adaptive learning, ai in education, ai-powered tutoring, curriculum design, digital literacy, edtech, educational assessments, educational chatbots, educational equity, educational policy, educational research, educational technology, education policy, inclusive education, lifelong learning, personalized education, student data, student privacy, teacher training, virtual classrooms
Environment_and_Sustainability	biodiversity, carbon sequestration, circular economy, climate modeling, climate resilience, conservation, eco-conscious, ecosystem services, energy efficiency, environmental impact, green infrastructure, ocean conservation, precision agriculture, renewable energy, sustainable agriculture
Ethics_General	accountability, ai ethics committee, ai governance, ai impact assessments, algorithmic bias, auditability, autonomy, reduce harm, beneficence, bias, civil society engagement, dual-use technology, ethical ai, ethical guidelines, explainability, fairness, fairness-aware algorithms, justice, non-maleficence, principles, privacy by design, public participation, responsible ai, stakeholder engagement, transparency, unintended consequences, value alignment, values

Existential_Risk	ai race dynamics, ai weaponization, ai-induced unemployment, artificial general intelligence, autonomous weapons, biological risk, chemical risk, control problem, critical, critical situation, endangering, endangerment of humanity, engineered pandemic, existential risk, future of humanity, global catastrophic risks, global catastrophe, human extinction, intelligence explosion, instrumental convergence, lethal autonomous weapons systems, maxipok, nuclear risk, orthogonality, peril, perilous, privacy erosion, superintelligence, threat, threat to humanity, threatening, unaligned artificial intelligence, value alignment problem, weapons of total destruction
Finance	algorithmic trading, anti-money laundering, blockchain in finance, credit scoring, financial compliance, financial consumer protection, financial ethics, financial inclusion, financial innovation, financial policy, financial regulations, financial risk management, financial stability, financial transparency, fintech, fraud detection, insurtech, regtech, responsible lending, robo-advisors
Healthcare	clinical decision support, diagnosis, electronic health records, health ai, health data interoperability, health informatics, healthcare ai adoption, healthcare ai governance, healthcare ai safety, healthcare ai transparency, healthcare data sharing, healthcare ethics, healthcare policy,

	healthcare regulations, medical imaging, patient privacy, personalized medicine, telehealth, telemedicine
Human_Rights_and_Social_Impact	digital rights, freedom of expression, human dignity, discrimination, non-discrimination, privacy, representation, right to access, right to autonomy, right to be forgotten, right to data protection, right to education, right to erasure, right to explanation, right to fair treatment, right to health, right to privacy, right to rectification, societal impacts of ai, social justice
International_Cooperation	cross-border collaboration, cross-cultural ai ethics, global ai collaboration, global ai governance, global ai impact assessment, global ai partnership, global ai policy forum, global ai research network, global norms, harmonization, intergovernmental cooperation, international ai conference, international standard, international treaty, multilateral agreement, transnational ai challenge
Regulation_and_Legal_Frameworks	ai regulations, compliance, enforcement, legal accountability, legal challenge, legal compliance, legal due diligence, legal frameworks, legal implications, legal liability, legislation, regulatory agility, regulatory audits, regulatory bodies, regulatory framework, regulatory harmonization, regulatory impact assessment, regulatory reporting, regulatory sandbox experiments, regulatory sandboxes

Research_and_Innovation	ai research, ai research ethics, collaboration, innovation incentives, interdisciplinary research, open science, research data sharing, research dissemination, research ethics boards, research ethics committees, research ethics guidelines, research ethics training, research funding, research impact, research infrastructure, research partnerships, research reproducibility, responsible ai innovation
Security_and_Cybersecurity	adversarial attack, cyber threat, cybersecurity awareness, cybersecurity framework, cybersecurity risk assessment, encryption, incident response, secure ai deployment, secure api, secure authentication, secure cloud computing, secure communication, secure data storage, secure development, secure supply chain, security audit, security patch, threat model, vulnerability
Stakeholder_Engagement	civil society engagement, industry consultation, multi-stakeholder dialogue, public participation, stakeholder accountability, stakeholder awareness, stakeholder collaboration, stakeholder diversity, stakeholder education, stakeholder empowerment, stakeholder feedback, stakeholder forum, stakeholder inclusion, stakeholder involvement, stakeholder perspective, stakeholder representation, stakeholder trust

Transportation	autonomous vehicles, connected vehicles, mobility as a service (maas), public transit, safety, smart cities, sustainable transportation, traffic management, transportation, transportation accessibility, transportation data, transportation emissions, transportation equity, transportation infrastructure, transportation regulations, transportation safety, urban planning