This is a preprint version of a chapter accepted for publication in the *Handbook on the Global Governance of Artificial Intelligence* (Edward Elgar Publishing, forthcoming 2026). Please cite the published version when available.

Future Trends in Global AI Governance

Daniel S. Schiff

Purdue University
https://orcid.org/0000-0002-4376-7303
dschiff@purdue.edu

Daniel S. Schiff is an Assistant Professor in the Department of Political Science and Co-Director of the Governance and Responsible AI Lab, Purdue University, United States

Abstract: This book chapter examines possible directions for future AI governance based on insights from early 21st century AI policy and discourse. It lays out a typology of key dimensions that serve as meaningful fault lines: decisions related to centralization versus decentralization, robust versus minimalistic monitoring and enforcement, adaptive versus enduring regulation, private leadership versus shared control, and a broad versus targeted scope of concern around AI's social and ethical impacts. After examining these characteristics and associated trade-offs, it reviews four analogous models in technology governance which bundle these dimensions in various ways: the aviation, climate change, organic food, and open-source software policy regimes. These models serve as a useful starting point for the chapter's examination of how AI governance is developing—including which paths appear more settled—and what key contingencies and opportunities remain. It concludes that persistent hard work will be necessary to shape a better future for all.

Keywords: AI governance, forecasting, policy design, technology policy, AI ethics

Word count: 11,035

1. Introduction

AI development and policy have been highly dynamic in the early 21st century as actors have continually competed to shape the frames, agendas, and institutions that will govern artificial intelligence (AI). While stakeholders agree on some core or general goals of a regulatory regime (such as the need to 'balance' innovation and social protection), strong disagreement remains on elements of key importance. Predicting the future of a highly dynamic set of technologies like AI is no easy feat; anticipating how the array of evolving governance institutions will adapt *in turn* is thus at least as challenging. However, while the shape of AI policy and governance may be contingent on the direction of AI development—amongst numerous other factors—there are tools we can employ to engage in policy forecasting.

In this chapter, I examine potential future pathways for how AI governance as a global project might unfold. The purpose of this exercise in policy forecasting is threefold: First, it aims to provide an analytic framework for reviewing key dimensions or debates along which AI governance might develop. Such a framework is subject to expansion or other updates, but can nevertheless serve as a useful baseline. Second, it enables us to take stock of recent history and assess where evidence may be pointing. This includes not only identifying where certain fault lines have begun to form, but recognizing what policy windows might persist or, indeed, open up over time. Third and relatedly, it can provide us with insight to chart our own future.

To do so, after introducing the chapter, Section 2 motivates and articulates a set of dimensions that increasingly characterize AI governance discourse (if not AI governance itself). These dimensions are presented as a (simplified) spectrum of possibilities with two poles: the tension between centralization and decentralization; the prospects for robust versus minimalistic monitoring and enforcement regimes; the likelihood of adaptive versus inflexible regulation; the debate between industry dominance or shared control; and demands for a broader or narrower scope of social response.

Yet these dimensions are not all orthogonal to one another. Various arrangements may be more likely. To provide some notional insight on future pathways for AI governance, Section 3 draws from historical and present-day models to present four possible speculative futures. I draw on the governance regimes surrounding aviation, climate change, organic food regulation, and open-source software, in each case describing both the appeals and limitations of such an approach for AI governance.

Next, Section 4 reviews key developments and trends in the first two decades of AI policy. In light of initial indicators of emerging governance models within individual countries and internationally, I consider which AI governance models appear most likely. Section 4 also takes into account the possibility of 'wildcards'

or critical junctures, noting opportunities for course correction and upcoming decisions that may have significant impact. Finally, Section 5 concludes with implications for future scholarship and policy.

2. A Typology of Key Dimensions in AI Governance

Governance as a concept refers to a broader, more inclusive notion of managing a particular issue, though it is related to more traditional notions of policy and regulation (Pierre & Peters, 2003). While it can include formal and mandatory laws and regulations, it can also incorporate voluntary standards, industry norms, the practices of individual companies and design teams, and even the practices adopted by individuals (Bullock et al., 2022; Shneiderman, 2022). That is, governance is a multi-layered, multi-stakeholder "collection of technologies and systems, people, policies, practices, and relationships that interact to support governing activities" (Johnston, 2010, p. 1).

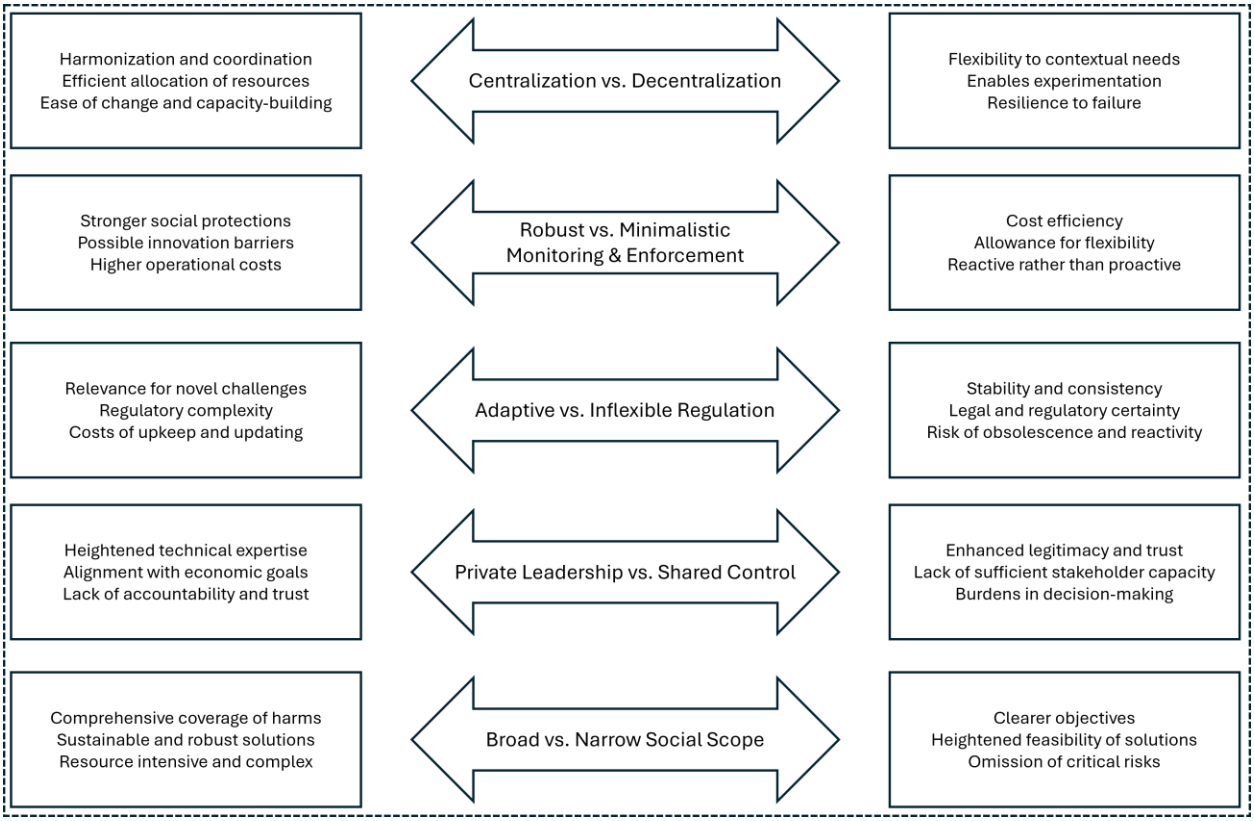

Figure 1. A typology: Key dimensions of AI governance and associated trade-offs

The contours and boundaries of governance regimes can thus be quite complicated, invoking everything from arrangements of actors to the design of institutions. However, as governance regimes differ by context (Fenwick & Vermeulen, 2018; Misuraca, 2009), it's possible to elicit particular dimensions that are of special importance for domains like AI. Figure 1 presents five dimensions in the AI governance discourse that have been prominent since the origins of AI policy as a field (circa 2016) and as part of early AI policy agenda-setting (Schiff, 2022). For each dimension, it discusses possible characteristics— including key potential upsides and downsides—of moving along a certain gradient. Note however, that the poles for each dimension are most safely interpreted as heuristic, covering only a subset of possible benefits and risks.

2.1 Centralization versus decentralization. A common and persistent question in AI governance is to what extent various institutions, regimes, and even practices should be designed, structured, or implemented out of a central (typically higher-level) authority. Note that this question applies to everything from the structure of international AI governance (e.g., should there be an international agency that coordinates research or monitors AI risks) to the structuring of employees within individual companies (e.g., should there be an AI ethicist on individual teams or a central team of AI auditors). Both innovation and protection from AI-related harm are impacted by this structural issue. There is much we know about centralization, as debates are recurring in public policy, management science, economics, healthcare, education, international relations, IT infrastructure, innovation systems, and other fields. Moreover, it is worth noting that preferences vis-à-vis centralization can shift over time, as in a pendulum swing (Evaristo et al., 2005; Steiner-Khamsi * & Stolpe, 2004), and can also serve as a proxy for other 'real' political debates.

A clear example of the possible variation that may unfold here can be seen with the AI governance strategies of the European Union (or China) compared to that of the United States (or UK or Brazil). For example, the European Union's AI Act creates a centralized AI Office authorized under a mandatory, horizontal regulation that applies to all European Union member states. Standards for permissible and impermissible usage of AI, definitions, documentation requirements, auditing and enforcement, and so on are explicitly designed to apply consistently in a centralized fashion to support the EU's goal of harmonization, protecting rights and advancing trade in the single market. In contrast, the prominent paradigm in US AI governance has been decentralized and vertical governance, with preference for sector specific design, implementation, and monitoring of AI (Díaz-Rodríguez et al., 2023; Park, 2024).

Cihon et al. (2020) provide an excellent review of associated trade-offs in the context of AI. For example, in addition to obvious advantages of concentrated authority and political influence that can be efficiently allocated, they note that centralized regimes could lower costs, create economies of scale, and even simplify participation, while also enhancing the prospects for coordination and policy foresight. However, decentralized institutions may be less brittle to failure or capture, could enable greater responsiveness and

flexibility to particularized needs, and could also facilitate easier local participation (Porter & Olsen, 1976). Note then that some trade-offs are not straightforward or outcomes remain unclear (Taylor, 2007). For example, while decentralized institutions could update more rapidly with respect to local needs, centralized institutions could enable more pervasive and rigorous reforms if the requisite capacity and negotiations are in place.

The centralization debate connects with numerous issues, like the novelty and dynamism of AI and AI governance, uncertainty around AI's impacts, challenges surrounding global coordination or strategic competition, prospects for forum shopping, stakeholder influence competition, and other aspects of dimensions in the framework. While much remains unknown and AI governance could develop anywhere along the spectrum (or in multiple places depending on the unit of analysis), this question is likely to remain absolutely essential.

2.2 Robust versus minimalistic monitoring and enforcement. A second important topic related to the development—and success—of regulatory regimes is whether monitoring enforcement are in place and well-designed. This is closely related to another issue, the debate between industry self-regulation and external formal regulation by government. However, for the sake of this typology, it is fruitful to examine what is often a key implicit entailment of the self-regulation versus formal regulation debate, which is the likelihood for rigorous monitoring and enforcement.

In short, effective monitoring and enforcement are often seen as essential to the overall effectiveness of policy, particularly regulatory policies. This is especially prevalent in contexts like environmental protection, food safety, healthcare, transportation and more (Gray & Shimshack, 2011; Harris & Raviv, 1978; Powell et al., 2013). Monitoring regimes can involve numerous activities at multiple scales and time periods, such as pre-deployment testing and validation, post-market monitoring, independent auditing, opportunities for whistleblowing and incident reporting, etc. Relatedly, monitoring regimes are often thought to be ineffective unless they are accompanied with sufficient enforcement in the form of incentives or punishments (Armour et al., 2020; Kambhu, 1989), even in the case of partial self-regulation (Ruhnka & Boerstler, 1998).

Unsurprisingly, there is substantial discussion around monitoring and enforcement regimes in AI policy, with proposals surrounding AI ethics and governance auditing (Birhane et al., 2024), safety testing (Anderljung et al., 2023), red teaming (Perez et al., 2022), assessment of large fines by actors like the EU, various other liability schemes (Buiten et al., 2023), and still other mechanisms. Some strategies target individual AI models or systems, others teams or companies, and still others seek to monitor activities of state actors. Like with questions of centralization, the devil is in the details.

However, while enhanced monitoring and enforcement can improve regulatory effectiveness if well-designed (a sizable challenge), it is not without risk. Developing such a regime is expensive and can introduce burdens that slow innovation, in the worst case without actually achieving the desired protections. In contrast, more minimalistic forms of enforcement might be sufficient to protect against harms, allowing for individual sectors or organizations to use existing and more familiar regulatory mechanisms (Short & Toffel, 2008). For example, allowing companies with greater technical expertise and access to the internal workings of their organizations to devise the most efficient strategies for ex-ante monitoring could be more feasible than treating a large cohort of expert government auditors or relying on ex-post liability (Innes, 2004). Again, the 'best' solution is likely to vary based on context and lie somewhere along the gradient.

2.3 Adaptive versus enduring regulation. A heightened problem in the context of technology governance is the pacing problem, where technology development is rapid enough that policymakers can only play 'catch up' (Marchant et al., 2011). Indeed, policy initiation, agenda setting, evaluation of alternatives, and implementation and evaluation can take many years to do, much less get right. This gap is exacerbated even more in the case of AI, where capabilities development has been rapid and societal impacts are likewise highly uncertain. Indeed, AI is already pervasive across a huge number of social sectors, leading to urgent calls to 'regulate now' (Jacobson et al., 2025), contested with criticisms that proposed regulation is poorly thought out or premature (Goldman, 2024; Gutierrez & Marchant, 2021). As argued in a letter by key industry lobbying groups, premature AI legislation may create "significant regulatory uncertainty by mandating compliance with novel requirements that rely on standards that are overbroad, vague, and impractical, if not infeasible" (Daylami, 2024, p. 2). As one example, an intensive debate surrounds even the definition of AI, with concern about specifying a too broad or narrow set of tasks and techniques.

At the same time, actors in industry have called repeatedly for definitional clarity and certainty, recognizing the benefits of harmonizing regulation to avoid having to comply with numerous inconsistent regulations (Kang, 2023). This would increase the pressure to come up with stable and consistent policies, including the definition of AI, which use cases are deemed high risk or prohibited, and so on. However too much stability can risk obsolescence and reactivity. Exemplifying this, the European Union's many actors who had been focusing intensively on AI for some years were largely blindsided by the popularization of generative AI in late 2022. This technological change shifted the definition and popular understanding of AI (drawing on the definition of the OECD), altered private sector and nation-state power dynamics, surfaced new concepts like 'frontier models' and policy solutions like red teaming (Fernández-Llorca et al., 2024; Liesenfeld & Dingemanse, 2024). Policymakers worked quickly to adapt their approach, but updates to major regulations are likely to become more difficult now that the AI Act is enacted.

Cognizant of this tension, policymakers have called for regulations to *both* provide certainty and allow for flexibility and adaptability. The United States National Institute of Standards and Technology (NIST) AI risk management framework articulates that "The AI RMF is designed to address new risks as they emerge. This flexibility is particularly important where impacts are not easily foreseeable and applications are evolving" (NIST, 2023, p. 4). The EU includes provisions (and evaluative criteria) for updating its list of prohibited and high-risk practices, as described in Articles 6 and 7 and elsewhere in the Act. For instance, Article 96 states that "At the request of the Member States or the AI Office, or on its own initiative, the Commission shall update guidelines previously adopted when deemed necessary" (Artificial Intelligence Act: Regulation (EU) 2024/1689, 2024, p. 114). Yet it is not clear how to strike such a balance. For instance, debates surrounding this dimension have now extended to concepts like general-purpose AI, where capabilities are improving regularly. One proposal is to identify important features like the number of parameters, size of a data set, and amount of computation, without providing fixed thresholds for each.

2.4 *Private leadership versus shared control*. If the first few dimensions relate to structure, rigor, and temporality, the next dimension relates to ownership. It considers who is best suited to lead AI governance. On the one hand, private sector actors retain major advantages, especially large multinational technology companies. These firms, not governments, are responsible for the majority of AI research and development, including much of the most cutting-edge frontier or foundational models. This includes decision-making around both innovation and governance strategies affecting individual AI systems and overall priorities (Khanal et al., 2024). Likewise, large private actors benefit from some of the strongest expertise, arguably both in the technical/computational and sociotechnical dimensions, owing to their size and ability to pay for talent (Ahmed & Wahed, 2020; Hopkins & Booth, 2021). Furthermore, these organizations have access to key internal information: trade secrets, details about data sources and model architecture, know-how on the feasibility of organizational structure and ethical design practices, and so on (Ubaydullaeva, 2024).

In addition to these advantages, private firms have natural incentives to cement their own leadership, both to secure favorable regulatory regimes and to secure first mover advantages against competitors or new entrants (Casanova, 2020; Makadok, 1998). Unsurprisingly, there has been major support for industry leadership, including in government agency strategic documents which de facto recognize the need for industry expertise in AI above public engagement (Schiff, 2023). Initiatives like the Industry Led Frontier Model Form (2023) and Partnership on AI reflect the belief that industry is best positioned to lead AI innovation—and even governance— achieving national strategic and economic goals.

Yet this position runs in stark contrast to calls for greater public participation, which take a number of forms. Stakeholders in academia, government, civil society, and industry have called for pluralism in the form of: public participation, interdisciplinary engagement, workforce diversification, active participation by civil

society, and engagement by vulnerable or impacted groups (Birhane et al., 2022; Buhmann & Fieseler, 2021). This shared control, at the level of policy or even AI system design, is thought to be crucial in helping steer AI innovation in alignment with public values. For instance the White House Blueprint for an AI Bill of Rights, (2022, p. 15) calls for direct engagement and states that "automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system." On this vision, shared control should enhance legitimacy, accountability, and trustworthiness of AI systems. However, it's worth noting that there are serious concerns about the feasibility of public engagement, the possibilities of 'participation-washing,' and a lack of AI literacy amongst non-experts (Coeckelbergh, 2024; DeCario & Etzioni, 2021; Sloane et al., 2022).

2.5 *Broad versus targeted social scope*. A final, if essential dimension, is the breadth of focus of AI governance efforts. In particular, AI regulations and ethical practices can focus on a subset of issues or on the larger constellation, with associated trade-offs. Perhaps the most prominent area of focus is on so-called 'technical AI ethics,' including issues of algorithmic bias, privacy, and explainability or interpretability. This focus owes itself in part to the early attention in the mid-2010s by academics and media on certain technical problems and fixes that were highly salient to the computing community (Angwin et al., 2016), leading to the creation of entire subfields (specialists, journals, conferences) focused on issues like algorithmic fairness (Ouchchy et al., 2020).

There are benefits to considering such a targeted scope. Some such ethical aims can be (arguably) more easily translated into software tools, evaluation criteria, and even statistical metrics (Bellamy et al., 2019; Cihon, 2019; Luccioni et al., 2023). With clear objectives comes increased feasibility of solutions, which helps to explain why the emergent AI ethics auditing ecosystem has focused its tooling on these issues (Schiff et al., 2024). Monitoring and enforcement also become more feasible as organizations are no longer awash in navigating a sea of vague and politically contested set of aspirational goals.

The weakness of the targeted scope is clear then. It may omit many critical risks and harms, including harms that do not attach as easily to technical thresholds, to single AI systems, or to products created by individual companies (Havrda & Klocek, 2023; Human & Watkins, 2022; IEEE, 2019). Indirect, systemic, and 'qualitative' harms are likely to be treated far better with a broader scope in AI governance. Such a scope would be more comprehensive, covering human rights, human well-being, and the environment, harms to social and political cohesion, global inequality, and so on (Green, 2021; Owe & Baum, 2021; Stahl et al., 2023). Yet these comprehensive approaches are beset by criticisms of feasibility with beleaguered attempts to translate abstract concepts into operationalizable metrics. They require sociotechnical and qualitative evaluation, may be subject to gaming, and are potentially resource intensive to implement. This complexity

has spawned a small literature on the ethical principles-to-practices gap AI (Avin et al., 2021; Baxter et al., 2020; Corrêa & Santos, 2024; Schiff, Rakova, et al., 2021).

3. Possible Models for AI Governance: Aviation, Climate, Organic Food, and Open-Source Software

While the typology of AI governance dimensions and their trade-offs may be useful, it's important to reiterate that these dimensions are not orthogonal. Instead, they are likely to be highly entangled, such that certain combinations are more beneficial or functionally (or politically) feasible. This section presents four different governance models, all focusing on technological policy domains, characterizing both each regime and its fit with the conceptual framework. While the coverage of each regime is concise, it provides an initial look at possible directions for AI policy.

3.1 *The aviation model: Strict compliance and technocratic governance*. The aviation safety regime stands out as one of the most robustly enforced and well-harmonized settings, reflecting significant regulation and interdependence. The International Civil Aviation Organization (ICAO), an agency of the United Nations, adopt standards on infrastructure, safety testing, border crossing, air quality, accident investigation, and more (Abeyratne, 2014). Supported by an internal technical body of appointed experts, the Air Navigation Commission, the ICAO coordinates with over 190 member states and national authorities who enforce standards within local jurisdictions, like the US's FAA and EU's EASA. Decisions are thus largely— though not entirely—non-political, driven by expertise in aviation, safety, and engineering (Jönsson, 1981).

The regime dates back to the early 20th century, with ICAO's predecessor, the International Commission for Air Navigation (ICAN) operating until the mid-1940s when the "Chicago Convention," or the Convention on International Civil Aviation, was adopted. The ICAO continues to make regular updates to policy, including addressing issues such as drug regulation and environmental impacts (Bows et al., 2009), though it has some limitations due to its authorizing statutes. Finally, monitoring is extensive, including the Universal Safety Oversight Audit Programme (USOAP), though the quality of implementation and enforcement varies by country (Eilstrup-Sangiovanni, 2022).

In line with the conceptual framework, we can characterize the aviation governance regime as having: 1) high levels of central control but also devolution to decentralized actors for implementation; 2) robust and costly if imperfect monitoring and enforcement, for example, given resource gaps between high and low-income countries; 3) a largely enduring regulatory base but with efforts to adapt to new developments; 4) technocratic control, but led by the government with substantial private sector involvement; and 5) a fairly

targeted social scope, focused primarily on safety, rather than other social issues like environmental impacts, labor quality, or ethical supply chains.

3.2 *The climate change model: Multi-stakeholder and non-binding governance*. Efforts to manage concerning impacts on the climate, and associated environmental goals, are marked by a highly complex and multilayered governance ecosystem. International conferences and resulting treaties like the Kyoto Protocol in 1997 and the more recent 2015 Paris Agreement have taken a much more targeted-but-flexible approach, typically focusing on the target of emission reduction and global temperature stabilization, and either setting binding targets (for Kyoto) or allowing countries to set their own targets (for Paris). This strategy, in alignment with the United Nations Framework Convention on Climate Change (UNFCCC) has been taken to maximize participation, including to address historical debates about equity and responsibility and the capacity of low income countries to balance climate mitigation with economic development (Seo, 2017). Along these lines, countries can determine their own development priorities, transition to renewable energy, and approach to policies like carbon taxes or command-and-control regulation. Thus there is some breadth in scope beyond emissions reduction, including policies like technology transfer or equitable financing.

However, despite efforts to allow countries to set their own Nationally Determined Contributions (NDCs) and establish locally feasible policy, most countries have failed to meet their targets given the lack of enforcement mechanisms and the ability of countries to even withdraw from the agreements (Widerberg & Pattberg, 2017). As a result, there is significant variation in how countries and even subnational governments approach climate change governance, creating a loosely-coupled regime (Keohane & Victor, 2011). Notably though, numerous stakeholders are engaged in climate conversations, ranging from governments to private sector actors to highly active civil society groups, public activists and academics (Cadman et al., 2017).

We can characterize the climate change regime as 1) significantly decentralized; with 2) minimalistic enforcement if greater monitoring (as evidenced by the Intergovernmental Panel on Climate Change); 3) largely adaptive regulation with changing targets, technological practices, and political contexts; 4) marked by shared control; and 5) with a mixed breadth of scope with both core targeted elements (like emission reductions) but also broader social aspirations.

3.3 *The organic food model: Third-party certification with unstandardized labeling*. Organic and genetically-modified food regulation, focused primarily on certification and labeling, has largely developed to meet consumer preferences. While some government actors like the EU have developed more uniform standards, such as the EU Organic Logo, or the US's Organic Certification, there remains substantial debate

and lobbying by interested actors (Sønderskov & Daugbjerg, 2011; Wohlers, 2013). This involves providing labels or informal labels that may not convey meaningful or accurate information to consumers (e.g., "free range chickens"), lobbying to exclude certain producers (e.g., of soy milk or almond milk or plant-based foods) from the symbolic benefits of labels, or aiming to dilute labels so standards are easily met (e.g., allowing for significant pesticide use with "organic" foods) (Daugbjerg et al., 2014; Giannakas, 2002; Janssen & Hamm, 2012).

The organic food policy regime is largely decentralized, with primarily national governance, and with governance altogether absent in many countries. Distinctions between more precautionary and more innovation friendly regions are evident and connected with national and domestic economic interests. Further, much enforcement and monitoring is carried out by third-party auditing organizations, including some that are solicited by the private sector auditees themselves (Albersmeier et al., 2009; Kononets et al., 2023). This status quo has allowed producers to select either favorable labeling practices or favorable auditors. Overall, enforcement is limited and focused on reactive monitoring. However, this regime is still new and may evolve in the future.

We can characterize the organic food governance regime as 1) largely decentralized, though with a few centralized elements like standard logos; 2) presently minimalistic monitoring and inconsistent enforcement despite the growth of an auditing industry; 3) still adapting in line with food production practices and consumer concerns; 4) dominated by the private sector with some government baseline standards; 5) and moderately broad in scope, with some elements focused on animal welfare, human health and safety, and environmental sustainability.

3.4 *The open-source software model: Collaborative and transparent governance*. The most 'distinct' model analyzed here may be the open-source software (OSS) regime. Dating to the days of the early Internet, a swathe of stakeholders pushed for a decentralized and collaborative approach, which came to characterize initiatives like Linux (1991), Wikipedia (2001), GitHub (2008), and Stack Overflow (2008). The Open Source Initiative distinguished OSS from free software, laying out a set of 10 principles and reviewing licenses to mainstream their preferred concepts and enable this ecosystem (Tozzi, 2017).

This general philosophy permeates much of the governance structure of this loose regime (or indeed, social movement) (Scacchi, 2007). Projects, data, and code are typically shared openly, with development often carried out by potentially large groups of collaborators who use open-source licenses. Code reuse and adaptation is encouraged through explicit technical governance structures, which may enable some forms of monitoring and transparency. Even leadership may emerge organically based on interest and skill set, though perhaps unsurprisingly given the subject matter, leaders are typically technical experts. There are

some downsides of this model, like issues with recruitment and retention of unpaid contributors (Lee, 2006), homogeneity (in education or other sociodemographic factors) of contributors, and disagreements in multi-stakeholder projects. Thus, while some optimistic advocates have encouraged the use of OSS community values in domains like AI, including the potential of decentralized strategies to manage complex problems (Schweik, 2003), there are challenges related to technical expertise, inclusion, coordination, and sustainability. Finally, the movement has centered a range of social goals like transparency, privacy, equity of access, and even fair labor.

Overall then, the OSS policy regime is characterized by 1) its unusually highly decentralized and collaborative nature, with emergent and individual-level leadership in many cases; 2) ecosystem level transparency through community norms but no formal enforcement; 3) a largely adaptive approach to governance through best practices; 4) highly distributed leadership; and 5) a reasonably broad scope of social and ethical concern. Table 1 summarizes how each policy regime fits into the governance dimensions discussed in the conceptual framework.

Governance Dimension	Aviation	Climate Change	Organic Food	Open-Source Software
Centralization vs Decentralization	Highly centralized but with local implementation	Decentralized with shared responsibility across stakeholders	Largely decentralized with some central elements (e.g., logos)	Highly decentralized and collaborative
Robust vs Minimalistic Monitoring and Enforcement	Robust, well-established, costly monitoring and enforcement	Minimalistic enforcement, greater focus on monitoring	Minimalistic monitoring, inconsistent enforcement	Moderate, with technical transparency through open collaboration
Adaptive vs Enduring Regulation	Largely enduring, with occasional updates	Adaptive to evolving technologies and targets	Still adapting based on evolving practices and consumer preferences	Adaptive and flexible through open-source contributions
Private Leadership vs Shared Control	Government-led but with significant private sector involvement	Shared control across multiple actors including states, civil society, and industry	Dominated by the private sector with some public standards	Technical leadership with decentralized expertise
Broad vs Targeted Social Scope	Targeted scope, focused primarily on safety and technical compliance	Mixed scope: emissions focus with broader social aspirations	Moderate scope with some focus on animal welfare, human health, and environment	Broad scope: collaborative, often technical but with several social and ethical considerations

Table 1. Possible models for AI governance based on four examples from technology policy

4. Where Are We Headed in AI Governance? Expectations and Contingencies in the Early 21st Century

With an understanding of some key dimensions underlying debates and trade-offs in AI governance, and examples of four notional models in other spaces of science and technology governance, we are positioned to examine how the future of AI governance itself may unfold. This exercise in policy forecasting involves both projections based on the first decade of AI policy action and discourse and an acknowledgment of key contingencies that could shape the direction going forward.

4.1 *Competing approaches to centralization with weak international governance*. For some dimensions of the framework, the direction of travel is more clear, at least in the near term. With respect to centralization versus decentralization approach, we are beginning to see the regime emphasize national and sectoral control. While bilateral, multilateral, and even global AI governance efforts have been expounded on, dominant national economic interests and strategic economic and military contests mean that much AI governance emphasizes the well-being of the nation-state. For instance, international efforts led by the UN (e.g., the UNESCO Recommendation on the Ethics of AI), OECD (e.g., OECD AI Principles and Policy Observatory), the Global Partnership on AI, and Council of Europe (e.g., The Framework Convention on AI) focused primarily on principles-level consensus with additional non-binding recommendations (Council of Europe, 2024; UNESCO, 2021b). Thus while countries like the US, Brazil, Turkey, and EU member states are all signatories of the OECD's recommendations on AI (OECD, 2021), for example, national priorities, regulations, and implementation differ widely. The establishment of various national AI safety institutes reiterates the likelihood of national control, even if supported by international coordination.

Meanwhile, a debate is underway between the benefits of greater central control at the national or bloc level, perhaps with a centralized government agency focused on AI, against a sector-specific vertical governance regime. Two primary approaches have formed, suggesting a possible bipolar division in governance approaches. The former, centralized approach is more prominently visible in the EU and China (e.g., the EU's AI Act and AI Board and China's Next Generation Artificial Intelligence Development Plan), with the latter, decentralized approach more visible in the US (e.g., the voluntary NIST AI Risk Management Framework, White House Blueprint for an AI Bill Of Rights, and since-overturned Executive Order on Safe, Secure, and Trustworthy AI). In contrast, subnational and local actors are less often thought equipped with sufficient expertise and capacity to manage AI effectively compared to national actors. However, even in the United States, the establishment of the AI Safety Institute (since renamed the Center for AI Standards and Innovation) and federal attempts to ban the subnational regulation suggest some interest in governance consolidation.

Overall, the model is closest to the climate change regime. International coordination focuses on high-level shared goals, including ethical principles in the case of AI, while national or sectoral governance regimes have ample flexibility to determine the details. Nevertheless, it's important to observe that there are significant ongoing efforts related to developing truly international governance institutions (Maas, 2023). Some proposals approximate the approach of CERN (Conseil Européen pour la Recherche Nucléaire), emphasizing international cooperation on grand challenges, shared use of computational resources and data, and primarily driving research. Others focus on stricter forms of compliance and monitoring, emulating the IAEA (International Atomic Energy Agency). This strategy, often discussed in the context of advanced AI systems (Maas, 2023) would oversee compliance with treaties, aim to minimize proliferation of dangerous AI, and perhaps provide technical assistance or perform conformity assessments. Finally, another proposal for international governance most closely resembles the IPCC, where stakeholders across academia, civil society and government synthesize AI research, create consensus-based reports, and aim to provide an advisory service to guide global policy.

The last approach has seemed to gain the most traction, followed by shared research and finally truly international governance in the form of monitoring. The recurrent creation of scientific expert bodies and advisory panels in AI, especially in the EU, is also suggestive of this mindset. This reiterates the likelihood of a climate change-like regime for AI in the near term. Note that the challenges associated with the strategy have become increasingly clear: high-level targets with ample flexibility, minimal enforcement, and asymmetries in power that disfavor low-income countries, all of which may impede achievement of the overarching goals. Thus despite the majority of countries supporting restrictions on lethal autonomous weapons in UN and Council of Europe deliberations, for example, leading countries like the US and China have rebuffed efforts to bind their use of AI in the military (Csernatoni, 2024). Similarly, undermining the momentum gained in multiple international AI safety and standard summits, the US and UK criticized commitments made at the 2025 Paris AI Action Summit, indicating the brittleness of such voluntary international coordination.

4.2 *Experimentation in monitoring and enforcement with several pathways*. The emerging approach to monitoring and enforcement in AI resembles an amalgam of the models presented in the previous section. Some efforts—like in aviation or California's SB 1047—seek to emphasize strict testing, evaluation, verification, and validation of AI systems (Vassilev et al., 2022), with an emphasis on safety engineering as a science, and critically, large fines for non-compliance in jurisdictions like the EU. In parallel, organizations have advanced incident tracking databases (McGregor, 2021), echoing safety-critical sectors' use of standardized reporting. And as noted previously, much of the actual monitoring and enforcement encouraged by international principles and frameworks will be carried out based on national standards and

regulations, as in the climate change regime. For instance, the EU AI Act asks member states to designate national competent authorities responsible for market surveillance and conformity assessment.

Elements of the open-source software regime are also visible, with a tradition of open machine learning (code, models, data) currently under threat by intense industry competition and trade secrecy (Liesenfeld & Dingemanse, 2024). On this approach, private industry and government AI developers and deployers are encouraged to provide access to academic researchers and civil society, and to promote transparency through strategies like AI inventories, transparency reports, datasheets for datasets, and model cards or 'nutrition labels' (Pareek & Goncalves, 2024; White House, 2023, 2024). Community initiatives like crowd-sourced labeling of deepfakes and bias or cyber bounties are meant to encourage openness in (Globus-Harris et al., 2022). It is now typical for leading AI companies to provide lengthy transparency reports or safety evaluations, representing an interesting blend of voluntary accountability and emerging quasi-standards. The balance between aspirational open source norms and trade secrecy remains dynamic.

Meanwhile, a regime of auditing, certification, and labeling is under development with encouragement by regulation but leadership by private audit firms and AI ethics startups (Schiff et al., 2024) and accompanying allegations of ineffectual self-regulation or conflicts of interest by quasi-auditors who actually function as non-independent consultants (Birhane et al., 2024). Indeed, a substantial portion of AI governance is currently led by individual companies and voluntary initiatives like the Frontier Model Forum or Partnership on AI (Microsoft Corporate Blogs, 2023), and even encouraged in regulation that allows for voluntary compliance (such as the EU's Codes of Practice), leading to numerous allegations of ethics washing by industry (Bietti, 2020; Kroet, 2024). This state of affairs resembles governance of organic food. Nevertheless, if schema for auditing, labeling, and certification are well-designed, the institutions may be effective. The successful development of this approach may constitute a key contingency in AI governance. Indeed, any of the strategies described above (or combination of them) could be successful or unsuccessful depending on their design or capture by interested actors.

4.3 *Ongoing challenges in managing the rapid evolution of AI*. Some of the greatest contingencies in the future governance around the development of AI itself. AI as a field has infamously both suffered from and benefited from faulty predictions, hype, fear mongering, and technical optimism (Bewersdorff et al., 2023; Tez, 2016). Meanwhile, as discussed in Section 2, the vast majority of global policymaking efforts failed to account for the impacts of the democratization of generative AI in the early 2020s. In just a few years, the governance conversation expanded to encompass new terms (e.g., frontier models, multi-modal models, foundational models), 'new' risks (e.g., impacts on the artistic professions, intellectual property, CBRN dissemination, low-cost text-to-video deepfakes, etc.), and new policy solutions (e.g., required registration of model training runs above a certain threshold of compute, red teaming) (Anderljung et al., 2023). Despite

extensive efforts to catalog AI's social and ethical implications, including efforts to taxonomize hundreds of possible AI risks (Slattery et al., 2025), stakeholders were largely blindsided by generative AI.

On the one hand, this raises the real possibility that anticipatory governance will not be able to effectively account for dramatic changes brought on by the evolution of AI. Perhaps reliance on overarching ethical principles and some basic institutional design principles will provide useful infrastructure, but AI is strikingly dynamic even compared to the other technology policy regimes examined in this chapter. That is, because AI is arguably different from prior technologies, we may have the least to learn from precedent. In combination, we may not be well-positioned to create truly adaptive governance, nor governance that is effectively enduring, but rather forced to be reactive.

Yet, on the other hand, one could argue that global governance institutions did respond fairly rapidly to the advent of generative AI. This includes making significant changes to the EU AI Act in short order (despite intense bureaucratic inertia), launching several national AI safety institutes, and fostering new frameworks and standards focused on general-purpose AI (Barrett et al., 2024) amongst other initiatives. Perhaps the numerous actors involved in AI governance have indeed built some capacity for rapid ideation, framework development, and institution building. A remaining question is whether large-scale policy institutions and instruments (like national laws, agencies, standards, and auditing practices) will be able to evolve as quickly, or whether they will be marked by problematic path dependencies (March & Olsen, 1995). Dynamism could urge the need for creation of new responsive or anticipatory governance institutions, perhaps those that are more decentralized, led by organizational arrangements with less inertia, or empowered to experiment to keep up with AI.

4.4 *Private and expert dominance of AI despite calls for shared governance*. A dimension where the future development seems to have more clear fault lines is the contest between private and shared control. Since the beginning of the AI policy conversation, large countries (AI leaders like the US and China) and multinational technology companies have been criticized for exercising too much control, leaving smaller countries or SMEs aside. Yet, despite efforts to bring in a broader range of actors at the national level (AI4D, 2020; UNESCO, 2021a), the major AI players continue to have significant control over global discourse through the Brussels Effect (Siegmann & Anderljung, 2022). Meanwhile, large technology companies appear to have consolidated even greater power, with concerns that the resources needed to train competitive foundational models are prohibitive for all but the wealthiest companies (Ahmed & Wahed, 2020; Liesenfeld & Dingemanse, 2024).

Along these lines, calls for participation by the general public, vulnerable and marginalized groups, and civil society arguably are also unheeded. Despite almost ubiquitous agreements that diverse and public

participation in policy design and across the AI lifecycle is important (Buhmann & Fieseler, 2023; Coeckelbergh, 2024), and some evidence that the public can influence policymaker attention (Schiff, 2024), the role of the public remains mostly nominal (Schiff, 2023). Often, public participation is called for in the mission statements of documents but practical strategies consist of expert collaboration within governments and industry, with public participation limited to 'public comment' procedures dominated by experts.

There are reasons to expect the status quo to continue. AI is notoriously technically complex, even by the standards of other technology domains. AI has also not been highly salient to the public compared to other policy issues (Jacobson et al., 2025), though this could change. Experts in industry have heightened access to know-how, data, and the inner workings of organizational governance. Meanwhile, the public has struggled with AI literacy (DeCario & Etzioni, 2021). Thus, laudable efforts to build public participation through creative strategies (Buhmann & Fieseler, 2021; Setälä & Smith, 2018) may be unlikely to come to meaningfully shape AI governance unless there is a deeper sea-change in public participation and technology governance overall. That is, advancing shared control beyond government, academic, and (largely) industry expertise currently appears more aspirational, limited to the few countries, companies, and test projects that invest enough to advance public voice.

4.5 *Broader ambitions but a targeted focus on technical ethics and safety*. AI governance has largely centered around a set of key concerns and shared values (Fjeld et al., 2020; Jobin et al., 2019) despite the existence of a much broader array of social and ethical concerns (Schiff et al., 2021). Most notably, the focus on algorithmic bias, transparency (and explainability), privacy, and increasingly safety represent emphases on technical governance—strategies where reasonably straightforward software or organizational practices can theoretically achieve certain key goals. These governance foci, for example, are most likely to be centered in policy documents, standards, and auditing procedures (NIST, 2024; Schiff et al., 2024).

However, there is notable attention to a broader universe of risks, harms, and goals related to AI. Conversations surround misinformation, polarization, mental health, trust, human-computer interaction, fair compensation, hidden labor, exacerbation of inequality, meaningfulness at work, and more. Major documents like the White House Blueprint for an AI Bill of Rights and EU AI Act reflect recognition of the importance of sociotechnical evaluation, a broad approach to risk identification, and even incorporation of strategies like well-being or human rights impact assessments. Yet, operationalizing these aims is more difficult. Sociotechnical impact assessments often involve qualitative review or focus on process frameworks, as determination of specific technical outcomes or thresholds is difficult (if not impossible). Given the difficulty of translating these requirements into regulation, early regulations have thus emphasized clearer targets like auditing for AI bias (Automated Employment Decision Tools (AEDT),

2021). In turn, audit firms and their client companies have likewise emphasized these targets, even advancing governance through software-as-a-service tools.

In short, there is substantial momentum toward technical ethics, increasingly inclusive of AI safety practices. Bias, transparency, privacy, and safety (and other concepts that can benefit if imperfectly from narrow technical treatment) seem likely to remain the focus. Meanwhile, efforts to advance well-being, human rights, protect meaningful work, and so on, require a complex array of additional strategies. This may include targeted AI for good initiatives, civil society or media pressure, whistleblowing, agenda setting or shifting of public opinion (Bernhardt et al., 2021; Coeckelbergh, 2024; Human Rights Watch, 2021; Samoili et al., 2020), leading to an array of formal governance and informal strategies to address these harms. Simply, *it is much easier to govern algorithmic bias than it is to govern algorithmic justice.*

That does not mean that there will not be continued attention on the latter category of broader concerns, with meaningful progress, such as demands for fair compensation of artists whose data are used to train models, calls to address environmental impacts of AI, or calls to regulate social media companies that platform deepfakes. However, it does suggest that continued work and harder work is needed to address broader socio-ethical concerns. One result could be two layers of governance: 1) a more tightly coupled technical governance regime as with the aviation model; and 2) a larger and more diffuse set of social and policy reform goals contingent on the effectiveness of their advocates and feasibility of their solutions.

5. Conclusion

This chapter sought to lay out some of the key dimensions that have come to characterize AI governance in the early 21st century. Focusing on questions of centralization, monitoring and enforcement, the temporality of regulation, distribution of control, and the breadth of scope, it examined possible directions and trade-offs. Drawing on initial evidence from the first decade of AI policymaking and governance discourse, it suggested some dimensions where certain paths seem more likely (bipolar approach to centralization, dominance of private control, and a largely technical governance scope), and other trajectories where key decisions could shape the success or failure of AI governance. For example, the effectiveness of the monitoring and enforcement regime will depend on the success of standards organizations and auditing and certification schemes.

Still other contingencies or wildcards are worth noting. For example, the trajectory of AI development itself is very unclear. Progress could be steady, allowing for more gradual development of governance institutions, experimentation, and iteration, as AI models become increasingly more sophisticated and well adopted in society. Or progress could be marked by punctuations, like with the popularization of generative AI, or still by rapid acceleration that far outpaces our ability to adapt. The latter possibility is certainly the

subject of much concern (Anderljung et al., 2023; Bengio, 2024) but the technological frontier remains unknown.

Another possible scenario is of a particular focusing event, such as a catastrophic or highly popularized incident. This could look like electoral disruption due to deepfakes, or a major cyber incident or infrastructure attack, a collapse (or boon) in economic or labor conditions, or even a military incident involving autonomous weapons. In the policy literature, it is well-known that focusing events like airplane crashes, channeled through narratives and media attention, can significantly shape the direction of discourse and governance. Again, it is unknown which if any of these incidents may occur, but they could very well lead to major impacts. For instance, a rapid increase in awareness of deepfake incidents has arguably led to a quick proliferation of regulation of non-consensual intimate imagery, child sexual abuse material, and political deepfakes. Finally, global cooperation or contestation is likely to play an important role. Cooperation between the EU and US, or heightened (or lowered) tensions between the US and China could meaningfully change the prospects for global governance. Over the long-term, changes in the nature of global governance, such as heightened internationalism, isolationism, new modalities for public participation, expanded or restricted power of large technology companies, and more, could all change the set of possible pathways for AI governance.

This list of contingencies is incomplete and expanding on it conceptually and empirically is left as an exercise for future researchers. Additionally, researchers and practitioners may wish to focus on developing adaptive and responsive governance frameworks, deliberately experimenting with respect to different levels of centralization, AI auditing practices, participatory frameworks, and more. This review has shown that there is much to learn from analogous regimes of technology governance, despite many of the novelties of AI policy. And while some paths seem more likely than others, significant uncertainty remains. There is thus ample opportunity to avoid visible pitfalls and design governance institutions that will promote a better future.

References

Abeyratne, R. (2014). *Convention on International Civil Aviation: A Commentary*. Springer International Publishing. https://doi.org/10.1007/978-3-319-00068-8

Ahmed, N., & Wahed, M. (2020). *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research* (arXiv:2010.15581). arXiv. https://doi.org/10.48550/arXiv.2010.15581

AI4D. (2020). *About AI4D*. Artificial Intelligence for Development in Africa. http://ai4d.ai/about-ai4d/

Albersmeier, F., Schulze, H., Jahn, G., & Spiller, A. (2009). The reliability of third-party certification in the food chain: From checklists to risk-oriented auditing. *Food Control*, *20*(10), 927–935. https://doi.org/10.1016/j.foodcont.2009.01.010

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (arXiv:2307.03718). arXiv. http://arxiv.org/abs/2307.03718

Angwin, J., Larson, J., Mattu, S., Kirchner, L., & ProPublica. (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Armour, J., Gordon, J., & Min, G. (2020). Taking Compliance Seriously. *Yale Journal on Regulation*, *37*, 1.

Artificial Intelligence Act: Regulation (EU) 2024/1689, 2024/1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Automated Employment Decision Tools (AEDT), Pub. L. No. Local Law 144 (2021). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., Krawczuk, I., Krueger, D., Lebensold, J., Maharaj, T., & Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science*, *374*(6573), 1327–1329. https://doi.org/10.1126/science.abi7176

Barrett, A. M., Newman, J., Nonnecke, B., Hendrycks, D., Murphy, E. R., Jackson, K., & Madkour, N. (2024). *AI Risk Management-Standards Profile for General-Purpose AI (GPAI) and Foundation Models (V1.1)*. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/

Baxter, K., Schlesinger, Y., Aerni, S., Baker, L., Dawson, J., Kenthapadi, K., Kloumann, I., & Wallach, H. (2020). Bridging the Gap from AI Ethics Research to Practice. *FAT* `20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 682. https://doi.org/10.1145/3351095.3375680

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Bengio, Y. (2024). *International Scientific Report on the Safety of Advanced AI - Interim Report* (DSIT 2024/009).

Bernhardt, A., Kresge, L., & Suleiman, R. (2021). *Data and Algorithms at Work: The Case for Worker Technology Rights* (p. 46). UC Berkeley Labor Center.

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence*, *4*, 100143. https://doi.org/10.1016/j.caeai.2023.100143

Bietti, E. (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*, 210–219. https://doi.org/10.1145/3351095.3372860

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. https://doi.org/10.1145/3551624.3555290

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). *AI auditing: The Broken Bus on the Road to AI Accountability* (arXiv:2401.14462). arXiv. http://arxiv.org/abs/2401.14462

Bows, A., Anderson, K., & Mander, S. (2009). Aviation in turbulent times. *Technology Analysis & Strategic Management*, *21*(1), 17–37. https://doi.org/10.1080/09537320802557228

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. *Business Ethics Quarterly*, *33*(1), 146–179. https://doi.org/10.1017/beq.2021.42

Buiten, M., de Streel, A., & Peitz, M. (2023). The law and economics of AI liability. *Computer Law & Security Review*, *48*, 105794. https://doi.org/10.1016/j.clsr.2023.105794

Bullock, J. B., Chen, Y.-C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. M., & Zhang, B. (Eds.). (2022). *The Oxford Handbook of AI Governance* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.001.0001

Cadman, T., Maguire, R., & Sampford, C. J. G. (Eds.). (2017). *Governing the climate change regime: Institutional integrity and integrity systems*. Routledge.

Casanova, J. (2020). *Online Search Engine Competition with First-Mover Advantages, Potential Competition and a Competitive Fringe: Implications for Data Access Regulation and Antitrust* (SSRN Scholarly Paper 3647092). https://doi.org/10.2139/ssrn.3647092

Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Center for the Governance of AI, Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Cihon, P., Maas, M. M., & Kemp, L. (2020). Should Artificial Intelligence Governance be Centralised? Design Lessons from History. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–234. https://doi.org/10.1145/3375627.3375857

Coeckelbergh, M. (2024). Artificial intelligence, the common good, and the democratic deficit in AI governance. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00492-9

Corrêa, N., & Santos, J. (2024). Crossing the principle–practice gap in AI ethics with ethical problem-solving. *AI and Ethics*, 1–18. https://doi.org/10.1007/s43681-024-00469-8

Council of Europe. (2024). *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Council of Europe. https://rm.coe.int/1680afae3c

Csernatoni, R. (2024, July 17). Governing Military AI Amid a Geopolitical Minefield. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/research/2024/07/governing-military-ai-amid-a-geopolitical-minefield

Daugbjerg, C., Smed, S., Andersen, L. M., & Schvartzman, Y. (2014). Improving Eco-labelling as an Environmental Policy Instrument: Knowledge, Trust and Organic Consumption. *Journal of Environmental Policy & Planning*, *16*(4), 559–575. https://doi.org/10.1080/1523908X.2013.879038

Daylami, R. (2024, August 5). *RE Opposition to SB 1047*. https://ct3.blob.core.windows.net/23blobs/dde95778-18c4-452c-ab99-3f4346c5c2f8

DeCario, N., & Etzioni, O. (2021). *America needs AI literacy now*. Allen Institute for AI. https://pnw.ai/article/america-needs-ai-literacy-now/72515409

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896

Eilstrup-Sangiovanni, M. (2022). Ordering global governance complexes: The evolution of the governance complex for international civil aviation. *The Review of International Organizations*, *17*(2), 293–322. https://doi.org/10.1007/s11558-020-09411-z

Evaristo, J. R., Desouza, K. C., & Hollister, K. (2005). Centralization momentum: The pendulum swings back again. *Commun. ACM*, *48*(2), 66–71. https://doi.org/10.1145/1042091.1042092

Fenwick, M., & Vermeulen, E. P. M. (2018). *Technology and Corporate Governance: Blockchain, Crypto, and Artificial Intelligence* (SSRN Scholarly Paper 3263222). https://doi.org/10.2139/ssrn.3263222

Fernández-Llorca, D., Gómez, E., Sánchez, I., & Mazzini, G. (2024). An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09412-y

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, *2020–1*.

Giannakas, K. (2002). Information Asymmetries and Consumption Decisions in Organic Food Product Markets. *Canadian Journal of Agricultural Economics/Revue Canadienne d'agroeconomie*, *50*(1), 35–50. https://doi.org/10.1111/j.1744-7976.2002.tb00380.x

Globus-Harris, I., Kearns, M., & Roth, A. (2022). An Algorithmic Framework for Bias Bounties. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1106–1124. https://doi.org/10.1145/3531146.3533172

Goldman, S. (2024, July 15). It's AI's "Sharks vs. Jets"—Welcome to the fight over California's AI safety bill. *Fortune*. https://fortune.com/2024/07/15/california-ai-bill-sb-1047-fierce-debate-regulation-safety/

Gray, W. B., & Shimshack, J. P. (2011). The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence. *Review of Environmental Economics and Policy*, *5*(1), 3–24. https://doi.org/10.1093/reep/req017

Green, B. (2021). The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing*, *2*(3), 209–225. Journal of Social Computing. https://doi.org/10.23919/JSC.2021.0018

Gutierrez, C. I., & Marchant, G. (2021). *A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence*. Arizona State University Sandra Day O'Connor College of Law. https://lsi.asulaw.org/softlaw/the-report/

Harris, M., & Raviv, A. (1978). Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance, and Law Enforcement. *The American Economic Review*, *68*(1), 20–30.

Havrda, M., & Klocek, A. (2023). Well-Being Impact Assessment of Artificial Intelligence—A Search for Causality and Proposal for an Open Platform for Well-Being Impact Assessment of AI Systems. *Evaluation and Program Planning*, 102294. https://doi.org/10.1016/j.evalprogplan.2023.102294

Hopkins, A., & Booth, S. (2021). Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. *AIES 2021 FIX*, 12.

Human Rights Watch. (2021). *How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net* (p. 28). Human Rights Watch. https://www.hrw.org/sites/default/files/media_2021/11/202111hrw_eu_ai_regulation_qa_0.pdf

Human, S., & Watkins, R. (2022). *Needs and Artificial Intelligence* (arXiv:2203.03715). arXiv. https://doi.org/10.48550/arXiv.2203.03715

IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition* (p. 294). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

Innes, R. (2004). Enforcement costs, optimal sanctions, and the choice between ex-post liability and ex-ante regulation. *International Review of Law and Economics*, *24*(1), 29–48. https://doi.org/10.1016/j.irle.2004.03.003

Jacobson, R., Schiff, D. S., & Schiff, K. J. (2025). *The Emergence of Partisan Politics in AI Policy: Coalition Formation in a Nascent Policy Subsystem*.

Janssen, M., & Hamm, U. (2012). Product labelling in the market for organic food: Consumer preferences and willingness-to-pay for different organic certification logos. *Food Quality and Preference*, *25*(1), 9–22. https://doi.org/10.1016/j.foodqual.2011.12.004

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnston, E. (2010). Governance Infrastructures in 2020. *Public Administration Review*, *70*(s1), s122–s128. https://doi.org/10.1111/j.1540-6210.2010.02254.x

Jönsson, C. (1981). Sphere of flying: The politics of international aviation. *International Organization*, *35*(2), 273–302. https://doi.org/10.1017/S0020818300032446

Kambhu, J. (1989). Regulatory Standards, Noncompliance and Enforcement. *Journal of Regulatory Economics*, *1*(2), 103–114. https://doi.org/10.1007/BF00140020

Kang, C. (2023, May 16). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

Keohane, R. O., & Victor, D. G. (2011). The Regime Complex for Climate Change. *Perspectives on Politics*, *9*(1), 7–23. https://doi.org/10.1017/S1537592710004068

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society*, puae012. https://doi.org/10.1093/polsoc/puae012

Kononets, Y., Konvalina, P., Bartos, P., & Smetana, P. (2023). The evolution of organic food certification. *Frontiers in Sustainable Food Systems*, *7*. https://doi.org/10.3389/fsufs.2023.1167017

Kroet, C. (2024, September 30). European Commission appoints 13 experts to draft AI Code. *Euronews*. https://www.euronews.com/next/2024/09/30/european-commission-appoints-13-experts-to-draft-ai-code

Lee, J.-A. (2006). New Perspectives on Public Goods Production: Policy Implications of Open Source Software. *Vanderbilt Journal of Entertainment and Technology Law*, *9*, 45.

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787. https://doi.org/10.1145/3630106.3659005

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). *Stable Bias: Analyzing Societal Representations in Diffusion Models* (arXiv:2303.11408). arXiv. http://arxiv.org/abs/2303.11408

Maas, M. (2023). International AI Institutions: A Literature Review of Models, Examples, and Proposals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4579773

Maas, M. M. (2023). *Advanced AI Governance: A Literature Review of Problems, Options, and Proposals* (SSRN Scholarly Paper 4629460). https://doi.org/10.2139/ssrn.4629460

Makadok, R. (1998). Can first-mover and early-mover advantages be sustained in an industry with low barriers to entry/imitation? *Strategic Management Journal*, *19*(7), 683–696. https://doi.org/10.1002/(SICI)1097-0266(199807)19:7<683::AID-SMJ965>3.0.CO;2-T

March, J. G., & Olsen, J. P. (1995). *Democratic governance*. Free Press.

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. Springer Science & Business Media.

McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), Article 17. https://doi.org/10.1609/aaai.v35i17.17817

Microsoft Corporate Blogs. (2023, July 26). *Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum*. Microsoft. https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/

Misuraca, G. (2009). Futuring e-government: Governance and policy implications for designing an ICT-enabled knowledge society. *Proceedings of the 3rd International Conference on Theory and Practice of Electronic Governance*, 83–90. https://doi.org/10.1145/1693042.1693060

NIST. (2023). *AI 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

NIST. (2024). *The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals*. NIST. https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf

OECD. (2021). *State of implementation of the OECD AI principles: Insights from national AI policies* (311; p. 93). OECD. https://doi.org/10.1787/1cd40c44-en

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5

Owe, A., & Baum, S. D. (2021). Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00065-0

Pareek, S., & Goncalves, J. (2024). Peer–supplied credibility labels as an online misinformation intervention. *International Journal of Human-Computer Studies*, 103276. https://doi.org/10.1016/j.ijhcs.2024.103276

Park, S. (2024). *Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework* (arXiv:2303.11196). arXiv. https://doi.org/10.48550/arXiv.2303.11196

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red Teaming Language Models with Language Models* (arXiv:2202.03286). arXiv. https://doi.org/10.48550/arXiv.2202.03286

Pierre, J., & Peters, B. G. (2003). *Governance, politics and the state* (1. publ., 2. pr). Macmillan.

Porter, D. O., & Olsen, E. A. (1976). Some Critical Issues in Government Centralization and Decentralization. *Public Administration Review*, *36*(1), 72–84. https://doi.org/10.2307/974743

Powell, D. A., Erdozain, S., Dodd, C., Costa, R., Morley, K., & Chapman, B. J. (2013). Audits and inspections are never enough: A critique to enhance food safety. *Food Control*, *30*(2), 686–691. https://doi.org/10.1016/j.foodcont.2012.07.044

Ruhnka, J. C., & Boerstler, H. (1998). Governmental Incentives for Corporate Self Regulation. *Journal of Business Ethics*, *17*(3), 309–326. https://doi.org/10.1023/A:1005757628513

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., Delipetrev, B., European Commission, & Joint Research Centre. (2020). *AI watch: Defining artificial intelligence: towards an*

operational definition and taxonomy of artificial intelligence. (p. 97). Publications Office of the European Union. https://doi.org/10.2760/382730

Scacchi, W. (2007). Free/open source software development. *Proceedings of the 6th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*, 459–468. https://doi.org/10.1145/1287624.1287689

Schiff, D., Borenstein, J., Laas, K., & Biddle, J. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42. https://doi.org/10.1109/TTS.2021.3052127

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81–94. IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2021.3056286

Schiff, D. S. (2022). *Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy*. https://doi.org/10.17605/OSF.IO/KW8XD

Schiff, D. S. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy. *Review of Policy Research*, *40*(5), 729–756. https://doi.org/10.1111/ropr.12535

Schiff, D. S. (2024). Framing contestation and public influence on policymakers: Evidence from US artificial intelligence policy discourse. *Policy and Society*, *43*(3), 255–288. https://doi.org/10.1093/polsoc/puae007

Schiff, D. S., Kelley, S., & Ibáñez, J. C. (2024). *The Emergence of Artificial Intelligence Ethics Auditing*. *Big Data & Society*.

Schweik, C. (2003). The Institutional Design of Open Source Programming: Implications for Addressing Complex Public Policy and Management Problems. *First Monday*. https://doi.org/20230415213701000

Seo, S. N. (2017). Beyond the Paris Agreement: Climate change policy negotiations and future directions. *Regional Science Policy & Practice*, *9*(2), 121–141. https://doi.org/10.1111/rsp3.12090

Setälä, M., & Smith, G. (2018). Mini-publics and deliberative democracy. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, & M. E. Warren (Eds.), *Oxford handbook of deliberative democracy*. Oxford University Press.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Short, J. L., & Toffel, M. W. (2008). Coerced Confessions: Self-Policing in the Shadow of the Regulator. *The Journal of Law, Economics, and Organization*, *24*(1), 45–71. https://doi.org/10.1093/jleo/ewm039

Siegmann, C., & Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*. Centre for the Governance of AI. https://www.governance.ai/research-paper/brussels-effect-ai

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2025). *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence* (arXiv:2408.12622). arXiv. https://doi.org/10.48550/arXiv.2408.12622

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation Is not a Design Fix for Machine Learning. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–6. https://doi.org/10.1145/3551624.3555285

Sønderskov, K. M., & Daugbjerg, C. (2011). The state and consumer confidence in eco-labeling: Organic labeling in Denmark, Sweden, The United Kingdom and The United States. *Agriculture and Human Values*, *28*(4), 507–517. https://doi.org/10.1007/s10460-010-9295-5

Stahl, B. C., Leach, T., Oyeniji, O., & Ogoh, G. (2023). AI Policy as a Response to AI Ethics? Addressing Ethical Issues in the Development of AI Policies in North Africa. In D. O. Eke, K. Wakunuma, & S. Akintoye (Eds.), *Responsible AI in Africa: Challenges and Opportunities* (pp. 141–167). Springer International Publishing. https://doi.org/10.1007/978-3-031-08215-3_7

Steiner-Khamsi *, G., & Stolpe, I. (2004). Decentralization and recentralization reform in Mongolia: Tracing the swing of the pendulum. *Comparative Education*, *40*(1), 29–53. https://doi.org/10.1080/0305006042000184872

Taylor, M. Z. (2007). Political Decentralization and Technological Innovation: Testing the Innovative Advantages of Decentralized States. *Review of Policy Research*, *24*(3), 231–257. https://doi.org/10.1111/j.1541-1338.2007.00279.x

Tez, R.-M. (2016, December 15). Rocket AI: 2016's Most Notorious AI Launch and the Problem with AI Hype. *Medium*. https://medium.com/the-mission/rocket-ai-2016s-most-notorious-ai-launch-and-the-problem-with-ai-hype-d7908013f8c9

Tozzi, C. (2017). *For Fun and Profit: A History of the Free and Open Source Software Revolution*. MIT Press.

Ubaydullaeva, A. (2024). Know-How and Trade Secrets in Digital Business. *International Journal of Law and Policy*, *2*(3), Article 3. https://doi.org/10.59022/ijlp.162

UNESCO. (2021a). *Artificial intelligence needs assessment survey in Africa* (p. 86). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000375322

UNESCO. (2021b). *Draft Text of the Recommendation on the Ethics of Artificial Intelligence* (41 C/73; p. 39). https://unesdoc.unesco.org/ark:/48223/pf0000373434

Vassilev, A., Booth, H., & Souppaya, M. (2022). *Mitigating AI/ML Bias in Context: Establishing Practices for Testing, Evaluation, Verification, and Validation of AI Systems*. https://csrc.nist.gov/pubs/pd/2022/08/18/mitigating-ai-ml-bias-in-context/ipd

White House. (2022). *Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age* (p. 73). White House, Office of Science and Technology Policy. https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/

White House. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. White House, Office of Science and Technology Policy. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

White House. (2024). Federal AI Use Case Inventories. *AI.Gov*. https://ai.gov/ai-use-cases/

Widerberg, O., & Pattberg, P. (2017). Accountability Challenges in the Transnational Regime Complex for Climate Change. *Review of Policy Research*, *34*(1), 68–87. https://doi.org/10.1111/ropr.12217

Wohlers, A. E. (2013). Labeling of genetically modified food: Closer to reality in the United States? *Politics and the Life Sciences*, *32*(1), 73–84. https://doi.org/10.2990/32_1_73