# **The Unique Challenge of Governing AI: Lessons from the Digital Revolution**

*2196 words*

The rise of artificial intelligence (AI) is often described as the next great technological revolution, promising transformative impacts across virtually every sector of society. From self-driving cars and smart home assistants to clinical diagnostics and computational drug discovery, AI is rapidly becoming embedded into our daily lives and core economic activities.

On one hand, the potential benefits are staggering – AI could help solve humanity's greatest challenges, like curing disease, reversing climate change, and achieving abundant renewable energy. On the other hand, the risks are equally profound – AI systems that are unaligned with human values or operate in unintended ways could threaten individual privacy and freedom, economic and social stability, and even human autonomy and civilization if advanced AI surpasses human-level general intelligence.

Given these high stakes, governing the development and deployment of AI is arguably one of the most consequential challenges boards and policymakers will face in the coming decades. Is governing AI fundamentally different from previous technological disruptions that precipitated corporate governance and government oversight changes?

I argue that while there are certainly parallels that can be drawn from past periods of rapid technological change, the unique attributes of AI – including its general-purpose, scalable, and recursive self-improving nature – present unprecedented challenges that require novel governance approaches beyond simply extending existing governance frameworks. Just as the digital revolution prompted sweeping changes to business strategy and regulation, governing the "AI revolution" will necessitate a proactive, coordinated, and multi-stakeholder response at both the organizational and societal levels.

## The Digital Precedent: Emergence of the Platform Economy

To illustrate how disruptive technologies catalyze governance changes, we can look at the recent emergence of digital platforms and the platform economy enabled by the Internet, mobile computing, cloud services, and data analytics. Upstarts like Google, Amazon, Facebook, and Uber were able to scale online platform business models rapidly, capturing network effects to achieve dominance in search, e-commerce, social media, and mobility services, respectively.

For incumbent firms across many industries, digital platforms represented both a competitive threat from more agile, data-driven entrants and an opportunity to transform business models and access new markets and capabilities. Navigating this shift required boards and executives to enhance technological governance – assessing disruptive threats, developing digital strategies, investing in new skills and infrastructure, managing data and cybersecurity risks, and overseeing the ethical use of algorithms and AI systems underlying platform services.

Policymakers similarly had to grapple with the implications of data-centric platform models. Issues around user privacy, anticompetitive behavior, algorithmic bias, workers' rights, and tax avoidance spurred regulatory scrutiny and calls for updated laws governing the digital economy, including comprehensive data protection and AI governance frameworks.

While navigating technological disruption is never easy for organizations or governments, seasoned boards and policymakers have experience developing governance approaches for emerging technologies that initially present uncertainties. Typically, this involves:

1. Conducting ongoing horizon scanning to monitor technological developments
2. Developing robust systems for identifying and mitigating key risks
3. Establishing principles, policies, and oversight mechanisms
4. Building new internal capabilities through talent, training, and infrastructure investments
5. Engaging with external stakeholders to understand differing perspectives

Many of these elements have already been applied by forward-looking organizations to govern the adoption and use of AI systems within their operations and offerings. Indeed, AI governance frameworks like Ethics & Trust Principles, AI Risk Management frameworks, AI Ethics Boards, and Responsible AI Practices have proliferated in recent years.

## Why Governing AI is Uniquely Challenging

While governing AI shares some high-level commonalities with governing previous technological disruptions like the rise of digital platforms, there are key differences stemming from AI's distinctive nature as a general-purpose, scalable, and potentially recursively self-improving technology. These unique attributes make AI governance an especially formidable challenge that differs in kind, not just degree, from past governance hurdles.

### *General-Purpose Nature of AI*

Unlike many past transformative technologies with relatively specialized industrial applications, the core AI capabilities of machine learning, reasoning, and perception enable incredibly broad use cases spanning virtually every sector of society and the economy. Whereas governance for an earlier emerging technology could focus on specific domains like transportation, energy, or medicine, AI's general-purpose nature requires a holistic, cross-sectoral view, given its rapid permeation into every facet of human activity.

From autonomous weapons systems and cyber warfare to automated healthcare diagnostics, AI-driven financial services, intelligent transportation networks, and smart energy grid management, virtually no domain is unaffected by the disruptive potential of AI. This ubiquity poses unique challenges for policymakers and governance frameworks aiming to "get ahead" of AI's impacts across a vast frontier of applications and contexts.

### *Scalability & Network Effects*

The ability for AI software systems to be effortlessly copied and deployed at negligible marginal cost, combined with AI's tendency to exhibit increasing returns to scale via self-reinforcing feedback loops, means that both the beneficial and detrimental impacts can rapidly scale to global proportions given the right conditions. An insecure or misaligned AI system designed by one actor has the potential to inflict widespread harm across borders once released into the world.

Whereas governance regimes for past innovations often focused on local or regional containment, transformative AI systems' scalable and extraterritorial impacts necessitate coordinated global governance frameworks to manage their deployment and socio-economic reverberations. The porousness of modern networks amplifies both the difficulty of controlling malicious AI applications and the imperative for international cooperation on AI governance norms and standards.

### *Recursive Self-Improvement Potential*

Perhaps the most significant long-term AI governance challenge relates to the prospective ability of advanced AI systems to recursively redesign and improve their capabilities, potentially triggering an unpredictable "intelligence explosion" leading to superintelligent AI that radically exceeds human-level general intelligence. While the specifics of such a scenario are hotly debated, the existential risks of a superintelligence not robustly aligned with human ethics, goals, and values cannot be overstated.

This self-improving and self-modifying nature of advanced AI systems represents a phase change from previous paradigms of specified, constrained, and human-controlled technologies. It introduces unprecedented difficulties for conventional governance approaches that are more accustomed to overseeing narrowly scoped innovations. Safeguarding the development of superintelligent AI that remains compatible with human civilization may rank among the most formidable challenges humanity has ever faced.

### *Economic and Power Reordering*

AI is increasingly viewed by economists and experts as a general-purpose technology on par with past technological revolutions like the steam engine, electricity, and computing. This designation implies AI has the potential to fundamentally reshape entire economic sectors and labor markets by automating a vast array of tasks and production activities currently performed by human workers. The large-scale disruptions to employment and profound reallocations of capital and wealth enabled by AI could yield significant social instability and reordering economic and geopolitical power across firms, sectors, nations, and regions.

The high stakes of being a competitive leader or laggard in the AI revolution have already contributed to an "AI arms race" dynamic among major powers vying for strategic advantage. As governance frameworks must manage these destabilizing forces and incentive structures, participants will likely resist constraints threatening their relative positioning, amplifying the difficulty of aligning interests and instituting effective AI governance norms and mechanisms.

In summary, the unique combination of general applicability across virtually all domains, ease of scalability and borderless impacts, potential for radical runway effects beyond human control, and transformative economic and geopolitical consequences position the governance of AI as a fundamentally distinct challenge from past technological revolutions. While sharing common principles like risk assessment and multistakeholder input, new paradigms tailored explicitly for the AI reality appear inescapable, given what's at stake. Developing robust AI governance frameworks proactively is thus imperative for mitigating existential risks and realizing AI's tremendous beneficial potential for humanity's long-term flourishing.

## Towards New Approaches to Governing AI

Given the generality, scale, and high stakes of the AI governance challenge, I argue that fundamentally new governance approaches beyond incrementally extending current frameworks are critically needed. Specifically:

### *Multistakeholder Collaboration*

The widespread impacts of AI across virtually all sectors mean no single organization or jurisdiction acting alone has sufficient knowledge, capabilities, or authority to govern it comprehensively. As Dafoe (2018) highlights, the novel challenges of AI "may motivate new forms of multistakeholder cooperation between industry, civil society, and governments to complement existing governance mechanisms."

Effective AI governance requires deep and sustained collaboration across industry, academia, civil society, and governments - including innovative models of public-private partnership that bring together disparate stakeholder viewpoints and areas of expertise. As Scherer (2016) notes, "The major policy challenge will be to set up adequate governance structures that incorporate perspectives from developers, manufacturers, owners/operators, employees, consumers/customers, and ordinary citizens affected by AI systems."

### *AI-Specific Governance Frameworks*

Given the unique attributes of AI systems and their potential radical impacts, we cannot simply retrofit existing governance frameworks designed for previous disruptive innovations like digital platforms or biotechnology. As Russell (2019) argues, "Because of the mismatch between advanced AI systems and the legacy institutions that might try to govern them, we face unprecedented challenges of forethought and coordination."

Entirely new, AI-tailored governance frameworks, policies, regulations, and compliance mechanisms should be developed proactively through internationally coordinated efforts. While these can extend principles from existing frameworks like the GDPR's algorithmic transparency and accountability provisions, fundamentally novel governance constructs designed explicitly for governing advanced AI systems will likely be required.

### *Early-Stage Governance*

Whereas governance efforts have historically followed a reactive cycle - waiting for potentially harmful impacts to manifest before developing policy interventions - early-stage and proactive governance is critical for transformative AI, given the possibility of catastrophic risks materializing rapidly.

As Grace et al. (2018) highlight, based on a survey of AI experts, there are significant uncertainties around potential timelines for advanced AI development but a non-negligible chance of transformative AI occurring in the coming decades. The prospect of an advanced AI system sparking globally destabilizing disruptions necessitates early and concurrent governance efforts alongside AI research and development to maximize preparedness and minimize existential risks through "a globally coordinated approach to governing AI development" (Tuck, 2021).

## *International Coordination and Cooperation*

Since the benefits and risks of transformative AI systems are globally scalable and extraterritorial, internationally coordinated governance frameworks and cooperation mechanisms involving all major stakeholders are essential. Unilateral national policies are likely insufficient and could lead to misaligned incentives, regulatory arbitrage, and destabilizing inter-state conflict exacerbated by a dynamic akin to an "AI arms race" (Agrawal et al., 2018).

New multilateral institutions and/or multistakeholder governance bodies with enforcement powers may ultimately be required, going beyond the partnership models historically used for governing global commons areas like climate change or nuclear non-proliferation. As Scherer (2016) argues, "a high level of coordination between nations will be necessary to reach an effective global governance regime for AI systems."

## *Prioritizing AI Alignment*

In addition to governing the ethical design, development, and deployment processes for AI systems, a core strategic focus of AI governance frameworks should be on the technically challenging "AI alignment" problem - ensuring advanced AI systems are reliably aligned with human ethics, constitutional values, and long-term civilizational interests.

Technical AI alignment solution approaches centered on robustly encoding human preferences and developing motivated value learning systems and corresponding institutional control measures to ensure controlled, transparent development of transformative AI capabilities should be prioritized (Bryson, 2018). The challenges of aligning a future superintelligent AI system fundamentally reshape traditional governance priorities.

### *Dynamic and Adaptive Governance Governing*

Given the potential for advanced AI systems to radically reshape societies, economies, and institutions in difficult-to-predict ways, governing AI is unlikely to have a static equilibrium end-state. Instead, dynamic, iterative, and constantly evolving governance approaches - employing tools like techno-economic scenario planning - will be essential to proactively adapt policies and resource allocations in response to the transformative impacts of advanced AI capabilities (Agrawal et al., 2018).

As Russell (2019) articulates, robust governance for transformative AI will require "new tools for agility and adaptability, new processes for re-examining prior assumptions and revising policies accordingly, new institutions for coordinating stakeholder activity, and new approaches for instilling stable and beneficial goal structures as AI systems become increasingly capable over time."

## Looking Ahead

Navigating the societal transition to an AI-driven future may be one of humanity's greatest challenges – one with existential stakes that demand proactive and robust governance frameworks. While governing emerging and general-purpose technologies has always been difficult, the unique attributes of AI position it as an unprecedented governance challenge by many orders of magnitude compared to previous technological revolutions.

Merely extending existing governance approaches will likely prove insufficient for governing advanced AI systems with broad impacts, scalable and extraterritorial effects, and the radical prospect of recursive self-improvement beyond human-level intelligence. Instead, new models of multistakeholder collaboration, internationally coordinated AI-specific frameworks, proactive early-stage governance, and prioritizing AI value alignment are required, given the stakes involved.

Though formidable, the potential of AI to be a tremendously beneficial force for humanity that helps solve our greatest challenges provides strong incentives to get AI governance right. An undertaking of this importance and complexity necessitates coordinated efforts between industry, civil society, governments, and the international community. Fortunately, we have the opportunity foresight to begin developing robust governance approaches for transformative AI proactively – an advantage previous generations lacked in governing earlier technological revolutions. The time to lay the groundwork for governing advanced AI is now.

# References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press.

Bryson, J. (2018). AI alignment: why it's good to take the risk of potential misalignment seriously. AI & Ethics, 1-8.

Dafoe, A. (2018). AI Governance: A Research Agenda. Oxford: Governance of AI Program.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Scherer, M.U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harvard Journal of Law & Technology, 29, 353-400.

Tuck, R. (2021). AI for Good: A Robot Governance Charter. In Economic Policies for the Toward an AI World (pp. 177-205). Emerald Publishing Limited.