Governing AI With Intelligence

URS GASSER

Artificial intelligence is now decades in the making, but especially since the emergence of ChatGPT, policymakers and global publics have been focused on AI's promise and its considerable, even existential, risks. Driven by machine learning and other advanced computational methods, AI has become dramatically more capable. Benefits have already been realized in areas such as transportation, health care, and sustainable growth, and there are more to come. However, the benefits are matched by mounting concerns over safety, privacy, bias, accountability, and the spread of increasingly compelling misinformation created by generative AI. Lurking as well is the possibility that AI might outperform humans in some contexts, shrinking the sphere of human agency as more and more decisionmaking is left to computers.

While there is a growing consensus on the challenges of AI and the opportunities it offers, there is less agreement over exactly what sort of guardrails are needed. What instruments can we use to unlock the technology's promise while mitigating its risks? Across the globe, myriad initiatives attempt to steer AI in socially desirable directions. These approaches come in different shapes and sizes and include ethics principles, technical standards, and legislation. While no universal strategy is likely to emerge, certain patterns stand out amid the diversity—patterns that constitute a thickening web of AI norms. There are hints here as to what it might mean to govern this evolving technology intelligently.

A global quest for AI guardrails

Over the past few years, governments have been exploring and enacting national strategies for AI development, deployment, and usage in domains such as research, industrial policy, education, health care, and national security. While these plans reflect material priorities, they typically also acknowledge the need for responsible innovation, grounded in national values and universal rights. The responsibilities of AI developers may be articulated in regulatory frameworks or ethics guidelines, depending on the state's overall approach to technology governance.

In parallel to the enactment of national policies, private and public actors have crafted hundreds of AI ethics principles. There are commonalities among them—in particular, shared areas of concern—but also much nuanced distinction across geographies and cultures. The ethics push has been accompanied by standards-setting initiatives from organizations bearing acronyms like NIST (the US National Institute of Standards and Technology) and CEN (the Comité Européen de Normalisation, or European Committee for Standardization). Professional associations such as IEEE promulgate best practices. Meanwhile, legislative and regulatory projects aim to manage AI through "hard" rather than "soft" law. In the United States and Europe alone, hundreds of bills have been introduced at all levels of government, with the European Union's newly approved AI Act being the most comprehensive.

Now add in the efforts of international institutions. The Organisation for Economic Co-operation and Development, with its AI Principles, and UNESCO, with its Recommendation on the Ethics of Artificial Intelligence, have established normative baselines that inform national AI governance arrangements. The United Nations adopted its first resolution on AI, highlighting the respect, protection, and promotion of human rights in the design and use of AI. G7 and G20 countries are attempting to coordinate on basic safeguards. Among the most ambitious international projects is the Council of Europe's framework convention to protect human rights, democracy, and the rule of law in the age of AI.

More tropical rainforest than formal garden

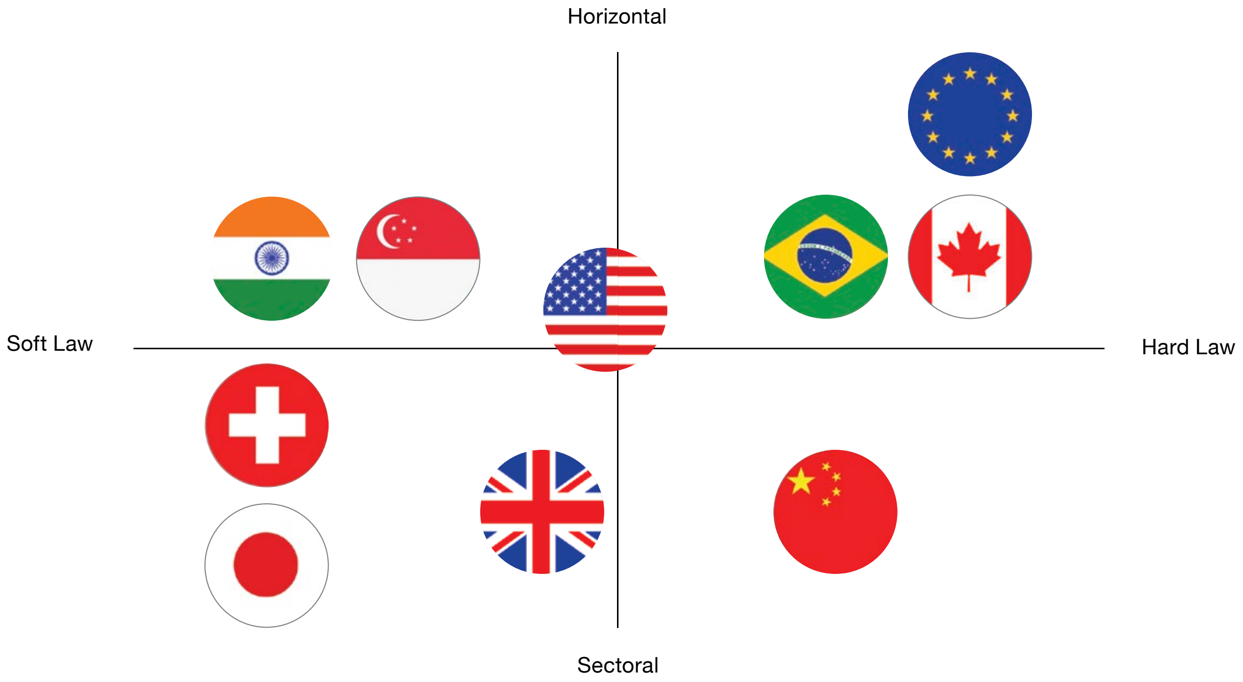
The landscape of AI governance, as these examples suggest, is no *jardin à la française*. Rather, it is as dense and intertwined as the Amazon. Importantly, from a governance perspective, the diversity isn't limited to the mere number of efforts underway. It is also reflected in the fact that governments and other rule-making institutions have pursued vastly different approaches.

Some countries, including Japan, Singapore, and India, rely to a large extent on the power of self-regulation and standard-setting to strike the balance between AI risks and opportunities. Canada, Brazil, and China, among others, take a more heavy-handed and government-led approach by enshrining rules guiding the development and use of AI in laws and regulations. Some jurisdictions are taking a "horizontal" approach by crafting rules intended to apply across most or all AI systems; others take more of a sector-specific approach, tailoring norms to industries and use cases.

One of the most comprehensive examples of the horizontal approach, targeting a wide range of AI applications, is the EU AI Act. Over dozens of pages, the law details requirements that developers and deployers of AI systems must meet before putting their products on the European market. The substantive and procedural requirements increase as one scales a pyramid of risks, with particularly strong safeguards for high-risk AI systems in sensitive areas such as critical infrastructure, education, criminal justice, and law enforcement. The AI Act, supplemented by sector-specific regulations, creates a complex oversight structure to monitor compliance and enforce rules by means of potentially hefty fines.

The United States has taken an alternative path. With gridlock in Congress, the Biden administration has issued the far-reaching Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. This order outlines a whole-of-government approach by establishing new standards for AI safety and security and

Figure 1. **NATIONAL APPROACHES TO AI GOVERNANCE**



launching programs across bureaucracies to safeguard privacy, advance equity and civil rights, protect consumers and workers, and promote innovation and competition. The initiative's hundreds of action items vindicate certain norms—for instance, against algorithmic discrimination—and, in general, aim to realize the high-level principles found in the White House's Blueprint for an AI Bill of Rights.

These and many other national and international governance initiatives form a complex and thickening canopy of principles, norms, processes, and institutions influencing the development and use of AI. The plurality of approaches, instruments, and actors involved, as well as the interactions among these, make AI governance a messy normative field that is increasingly difficult to navigate.

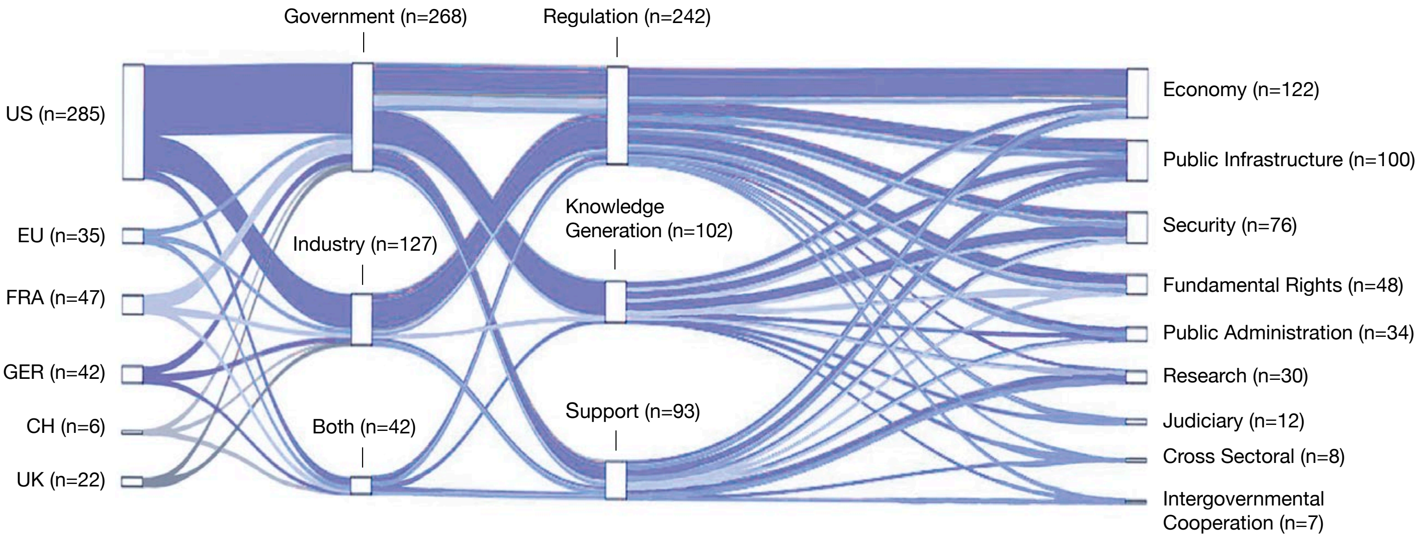
But while the rainforest teems with diversity at ground level, from a bird's-eye view, some functional patterns start to emerge.

Patterns in the landscape

One recognizable pattern across AI-governance arrangements relates to the different functions AI norms can play. Three kinds of norms are in operation today. First, and perhaps most intuitively, there are norms of constraint: AI norms typically place limits on the development and use of the technology. Such norms have been codified in rules such as bans on use of facial-recognition technology in some US cities and the premarket obligations for high-risk AI systems under the EU AI Act. A second category of norms, in contrast, is enabling. These norms permit or even promote the development and use of AI. Funding and subsidies reflect such norms. So do pro-innovation measures such as the creation of regulatory sandboxes—legal contexts in which private operators can lawfully test innovative ideas and products without following all regulations that might otherwise apply. Finally, a third category of norms attempts to create a level playing field. Such norms underlie, for example, transparency and disclosure obligations, which seek to bridge information gaps between tech companies and users and society at large; AI literacy programs in schools; and workforce training.

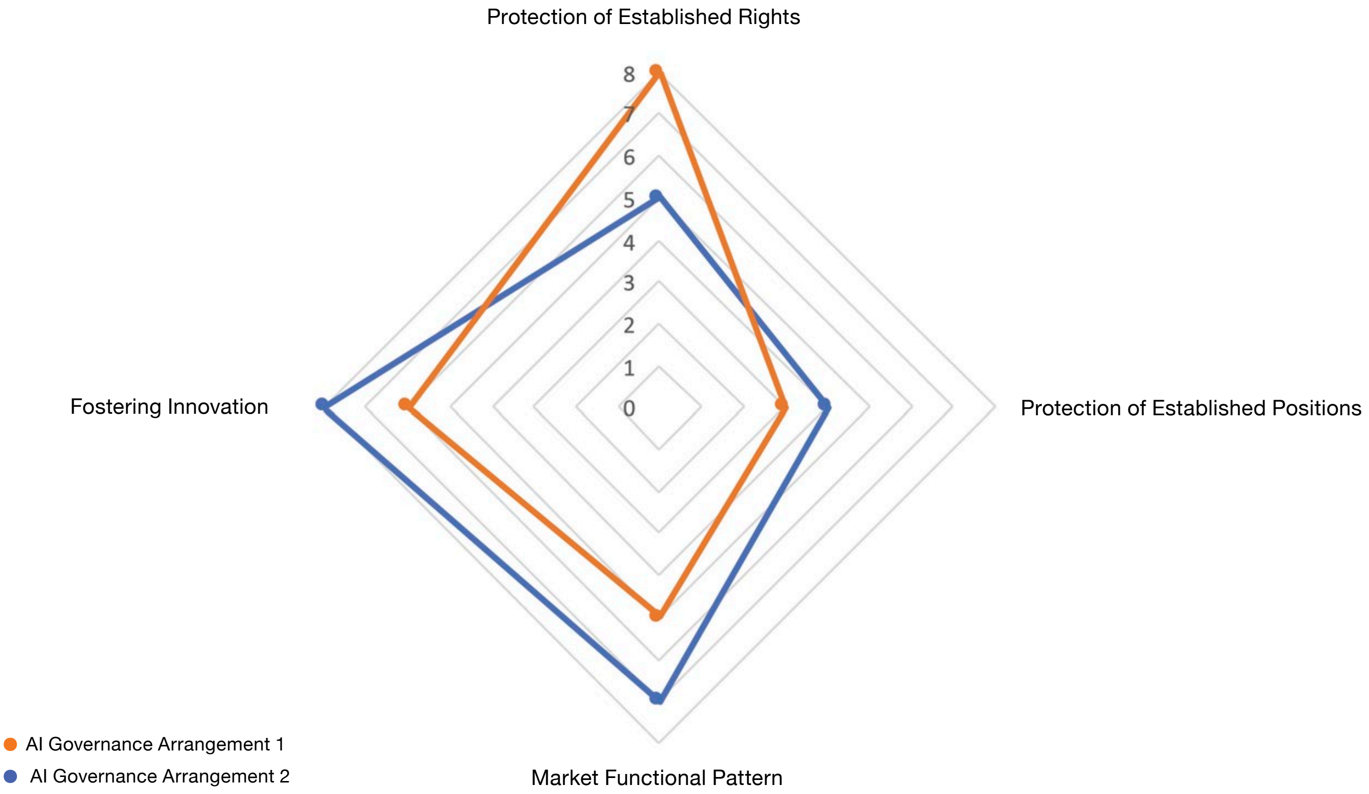
Assorted AI governance systems may emphasize constraining, enabling, and leveling norms to varying extents, but typically the aims of these systems are quite similar. That is, another pattern lies in the goals of diverse AI-governance arrangements, even as the means diverge. The protection of established rights is usually a top priority, so that governments use the norms available to them to shield citizens against discrimination, fraud, and privacy invasions that emerge from AI use. A related objective is protection of established interests, such as the economic goals of certain interest groups or industries. Many governance arrangements include rules designed to promote economic activity, stimulate technological development through market mechanisms, and support the creation of new markets and business models. More generally, innovation is a core theme of most AI arrangements, which often include provisions that seek to promote AI research and development.

Figure 2. AI REGULATIONS IN THE UNITED STATES AND EUROPE



US=United States, EU=European Union, FRA=France, GER=Germany, UK=United Kingdom, CH=Switzerland. Bills that included regulatory (constraining), knowledge generation (leveling) and/or supportive elements (enabling) were listed for each category separately. Source: Kerstin N. Vokinger, David Schneider, and Urs Gasser, "Mapping Legislative and Regulatory Dynamics of Artificial Intelligence in the US and Europe" (September 2023, manuscript under review).

Figure 3. NORMATIVE PATTERNS OF AI GOVERNANCE ARRANGEMENTS



Guardrails on the guardrails

The good news is that there is no shortage of approaches and instruments in the AI-governance toolbox. But, at the same time, policymakers, experts, and concerned members of the broader public cannot simply snap their fingers and see their governance goals realized. Contextual factors such as path dependencies, political economy, and geopolitical dynamics cannot help but shape the design and implementation of AI governance everywhere.

Whether in the United States, Europe, or China, the influence of national security interests on AI governance is becoming increasingly clear. The global powers are engaged in an AI arms race, with ramifications for the choices these leading states will make concerning the promotion of, and constraints upon, innovation. In particular, competitive dynamics dampen prospects for truly global governance—universally accepted and enforced AI rules. A case in point is the stalled discussion about a ban on lethal autonomous weapons systems.

Moreover, the norms, institutions, and processes constituting any approach to AI governance are deeply embedded in preexisting economic, social, and regulatory policies. They also carry cultural values and preferences in their DNA, limiting what is feasible within a given context. In other words, path dependencies caution against copying and pasting rules from one jurisdiction to another. For example, the EU AI Act cannot be transplanted wholesale into the laws of all countries.

Despite geopolitical tensions, economic competition, and national idiosyncrasies, there are islands of cooperation in the ocean of AI governance. International initiatives such as the UN AI Resolution, the G7 Hiroshima Process on Generative Artificial Intelligence, and the Global Partnership on AI seek to advance collaboration. These global efforts are supplemented by regional and bilateral ones, including, for instance, the transatlantic partnership facilitated by the US-EU Trade and Technology Council.

No formulas, but some insights

By this point, it is clear that there won't be a universal formula for AI governance any time soon. But I see three core insights emerging from a deeper analysis of the current state of affairs, which may inform initiatives in the near term and more distant future.

The first of these insights concerns learning. The rapid progress of technological development and AI adoption, combined with the lack of empirical evidence concerning what kind of governance interventions are likely to produce which outcomes, make AI a strong candidate for tentative governance. Tentative governance is a novel regulatory paradigm in which rules leave space for exploration and learning. In practical terms, whatever institutions take charge of AI governance—whether these are state institutions or industrial players—need to ensure that the rules they put forward are flexible enough to adjust in light of changing circumstances. It should be easy to update rules and eventually also to revise them heavily or revoke them when they no longer are fit for purpose. In addition, it is important to carve out spaces—think of controlled experiments—where certain guardrails can be lifted so that new AI applications can be tested. This is how all parties will find out about the risks of particular technologies and create ways to mitigate those risks. In short, learning mechanisms must be baked into AI governance arrangements because we often don't know enough about tomorrow to make lasting decisions today.

The second broad insight we might discern in today's fragmented AI-governance landscape is that great promise resides in interoperability among different regimes. Originally a technical concept, interoperability can be understood as the capacity of software applications, systems, or system components to work together on the basis of data exchange. But interoperability is also an advantageous design principle in the field of AI governance, as it allows for different arrangements—or, more likely, components of such arrangements—to work together without aiming for total unification or harmonization. Emerging AI-governance arrangements introduce and legitimize an assortment of tools and practices that may be thought of as modules subject to cross-border, multistakeholder cooperation. For instance, risk-assessment and human rights–evaluation tools could be aligned across otherwise-divergent AI-governance schemes.

The final insight speaks to capacity-building. Private- and public-sector actors who seek to develop or deploy AI systems in their respective contexts—health care, finance, transportation, and so on—are confronted with the challenge of translating high-level policies, abstract legal requirements, emerging best practices, and technical standards into real-life use cases. In order to support this translation, AI-governance initiatives should invest in implementation capacity, which includes AI literacy and technical assistance. Such capacity-building demands—once again—multistakeholder and increasingly cross-border cooperation and has significant implications for education systems. Experience with previous cycles of innovation suggests that these on-the-ground capacities are often as important as the policy choices made in halls of power. What's needed, ultimately, is governance in action, not only on the books.

Embracing opportunities for innovation—in both technology and governance

There can be little doubt that AI will have long-term effects on the inner workings of our societies. Right now, in universities and public- and private-sector laboratories alike, scientists and engineers with a zeal for innovation are creating new possibilities for AI, and public interest is high. There is little chance that this collective enthusiasm will abate any time soon.

But while technological innovation is propelled in whatever direction our desires and interests take, governance largely follows the narrow passage allowed by realpolitik, the dominant political economy, and the influence of particular political and industrial incumbents. We must begin to think beyond this narrowness, so that path dependencies do not overly constrain options for governance. This is a historic opportunity, a moment to engage fully and in a collaborative manner in the innovation not just of AI but also of AI governance so that we can regulate this transformative technology without squandering its potential. Traces of such innovation in current debates—outside-the-box proposals for new types of international AI institutions—should be recognized as invitations to embrace a worthwhile challenge: to design future-proofed guardrails for a world shaped by AI.

Despite geopolitical tensions, economic competition, and national idiosyncrasies, there are islands of cooperation in the ocean of AI governance.

***Urs Gasser** is professor of public policy, governance, and innovative technology at the Technical University of Munich, where he serves as dean of the TUM School of Social Sciences and Technology and rector of the Munich School of Politics and Public Policy.*

This essay is an edited version of a paper that was commissioned by the Annenberg Foundation Trust at Sunnylands.