A Layered Model for AI Governance

A Layered Model for AI Governance

Urs Gasser and Virgilio A.F. Almeida - Harvard University

AI-based systems are "black boxes," resulting in massive information asymmetries between the developers of such systems and consumers and policymakers. In order to bridge this information gap, this article proposes a conceptual framework for thinking about governance for AI.

Many sectors of society rapidly adopt digital technologies and big data, resulting in the quiet and often seamless integration of AI, autonomous systems, and algorithmic decision-making into billions of human lives[1][2]. AI and algorithmic systems already guide a vast array of decisions in both private and public sectors. For example, private global platforms, such as Google and Facebook, use AI-based filtering algorithms to control access to information. AI algorithms that control self-driving cars must decide on how to weigh the safety of passengers and pedestrians[3]. Various applications, including security and safety decision-making systems, rely heavily on A-based face recognition algorithms. And a recent study from Stanford University describes an AI algorithm that can deduce the sexuality of people on a dating site with up to 91 percent accuracy[4]. Voicing alarm at the capabilities of AI evidenced within this study, and as AI technologies move toward broader adoption, some voices in society have expressed concern about the unintended consequences and potential downsides of widespread use of these technologies.

To ensure transparency, accountability, and explainability for the AI ecosystem, our governments, civil society, the private sector, and academia must be at the table to discuss governance mechanisms that minimize the risks and possible downsides of AI and autonomous systems while harnessing the full potential of this technology[5]. Yet the process of designing a governance ecosystem for AI, autonomous systems, and algorithms is complex for several reasons. As researchers at the University of Oxford point out,3 separate regulation solutions for decision-making algorithms, AI, and robotics could misinterpret legal and ethical challenges as unrelated, which is no longer accurate in today's systems. Algorithms, hardware, software, and data are always part of AI and autonomous systems. To regulate ahead of time is difficult for any kind of industry. Although AI technologies are evolving rapidly, they are still in the development stages. A global AI governance system must be flexible enough to accommodate cultural differences and bridge gaps across different national legal systems. While there are many approaches we can take to design a governance structure for AI, one option is to take inspiration from the development and evolution of governance structures that act on the Internet environment. Thus, here we discuss different issues associated with governance of AI systems, and introduce a conceptual framework for thinking about governance for AI, autonomous systems, and algorithmic decision-making processes.

The Nature of AI

Although AI-based applications are increasingly adopted in hospitals, court rooms, schools, at home, and on the road to support (and in some instances, even guide) human decision-making, currently no universally accepted definition of AI, a term coined in the mid-1950s by US researchers[6][7]. One reason for the lack of a definition is that AI, from a technical perspective, is not a single technology, but rather a set of techniques and sub-disciplines ranging from areas such as speech recognition and computer vision to attention and memory, to name just a few[6].

From a phenomenological perspective, however, the term AI is often used as an umbrella term to refer to a certain degree of autonomy exhibited in advanced health diagnostic systems, next-generation digital tutors, self-driving cars, and other A-based applications share. Often, such applications in turn impact human behavior and evolve dynamically in ways that are at times unforeseen by the systems' designers. In this context, the differentiation between weak (or narrow) and strong (or general) AI is often used and helpful when discussing the nature of AI. Weak AI describes the current generation of applications that are focused on a relatively narrow task such as playing a game, recognizing a voice, or detecting certain patterns on a CT-scan. Strong AI, in contrast, refers to machines with genuine intelligence and self-awareness in the sense that the machine has the ability to apply intelligence to any problem[8]. At present, the technical possibility and (potential) societal impact of strong AI is discussed controversially, while the current adaptation of weak AI already leads to a series of real governance issues that deserve attention in the present.

AI Governance Challenges

Following a typical pattern when new technologies become more widely available, policymakers and other stakeholders are focusing largely on the risks and harms of AI-based technologies[9]. Again, similar to previous conversations about digital technologies' impact on society, the challenges related to AI, autonomous systems, and algorithms are often presented and discussed in the form of lists of substantive issues (including policy, legal, governance, and ethical considerations) that must be addressed[6].

A recent roadmap on AI policy by one leading expert, for instance, identifies the following clusters of core issues and questions where AI applications either lead to new challenges or amplify pre-existing policy concerns and pressure points[10]:

- *Justice and equality.* To what extent can AI systems be designed and operated to reflect human values such as fairness, accountability, and transparency and avoid (new) inequalities and biases?
- *Use of force.* As AI-based systems are now involved in making decisions about the use of force — for instance, in the case of autonomous weapons —

how much human control is necessary or required? Who bears responsibility for the AI-based outputs?

- *Safety and certification.* Particularly where AI-based systems have a physical manifestation, how do we define and validate safety thresholds — for instance, through standard-setting and certification?
- *Privacy.* As AI-systems are enabled and powered by data, what are the privacy implications and new privacy threats of next-generation technologies — for instance, in terms of government surveillance or corporate influence over customers?
- *Displacement of labor and taxation.* To what extent will AI-based machines replace jobs previously performed by humans, or at least transform what labor means? What are the effects of AI on public finances if robots don't pay taxes?

Such lists of substantive issues, several others could be added, (for instance, intellectual property or liability), can be supplemented by cross-cutting themes surrounding transparency, accountability, and explainability; inclusion and fairness; global governance; and more that span across the different application areas of AI-based systems (see).

Models for AI Governance

When considering future governance models for AI that address the aforementioned issues, it might be helpful and necessary to move beyond such lists and consider some of the larger structural challenges associated with the "regulation" (broadly defined) of AI-based technologies. In the following, we highlight three such challenges that translate into design requirements for a future governance model of AI.

Information asymmetries. While AI has the potential to shape the lives of billions of people, only a few experts really understand the underlying techniques. AI-based systems are often inscrutable, sometimes resulting in massive information asymmetries between the developers of such systems and other stakeholders, including consumers and policymakers. An effective governance system for AI needs to incorporate mechanisms aimed at improving our collective understanding of the AI phenomenon in its different manifestations and contexts of application.

Finding normative consensus. The current policy and governance debate is largely focused on risks and challenges associated with AI. But AI also offers tremendous potential benefits to society, as the discussions about the use of AI in the context of Sustainable Development Goals illustrate (see). A governance model must open up spaces for cost-benefit analyses and normative consensus building among different stakeholders, particularly where tradeoffs are involved in the design of AI systems. A future governance model also needs to deal with normative differences among contexts and geographies, and provide for

interoperability among different frameworks and approaches[11].

Government mismatches. Even where we have a shared understanding of AI technologies, the underlying techniques, and societal consensus about what is or isn't desirable, the design of effective, efficient, and legitimate means (strategies, approaches, tools, and so forth) to resolve the aforementioned substantive issues is challenging, given the conditions of uncertainty and complexity in the AI ecosystem. But larger undercurrents also put limits on traditional approaches to law- and policymaking in the digital age[12].

Taken together, these structural challenges and associated design requirements for a future governance model of AI point away from simple state-centric, command-and-control regulatory schemes toward more complex approaches to governance emergent in fields as diverse as the Internet, nanotechnology governance, or gene driver governance. While the exact contours of a future AI governance model are still in flux, advanced governance models such as active matrix theory, polycentric governance, hybrid regulation, and mesh regulation can provide both inspiration and conceptual guidance on how such a future governance regime might be designed[13]. In the next section, we highlight one feature that is common across many of these models: the idea of modularity embodied in the form of layered governance, which also combines different instruments to grapple with and address the aforementioned substantive issues, making it a shared responsibility among all relevant actors involved. It is important to note that any such emerging model must be situated in and interact with existing institutional frameworks of applicable laws and policies, particularly human rights, as the development and deployment of AI does not take place in a vacuum[14].

The Layered Model

Modularity is one of the main mechanisms for managing complex systems. Modularity aims to reduce the number of interdependencies that must be analyzed by identifying which tasks are highly interdependent and which ones are not[15]. Layering represents a particular form of modularity, in which different parts of the overall system are arranged into parallel hierarchies. A frequently cited example of layering is the Open System Interconnection (OSI) Reference model during the late 1970s[15]. Another example of a layered model was proposed by David Clark[16] to represent the nature of cyberspace using a model with four layers, that are: first, the people who participate in the cyber-experience; second, the information that is stored, transmitted, and transformed in cyberspace; third, the logical building blocks that make up the services, and fourth, the physical foundations that support the logical elements. The scale, heterogeneity, complexity, and degree of technological autonomy of AI systems require new thinking about policy, law, and regulation. We attempt to capture the complex nature of AI governance by using an analytical model with three layers. From the top down, the interacting layers are as follows:

- social and legal;
- ethical; and
- technical foundations that support the ethical and social layers.

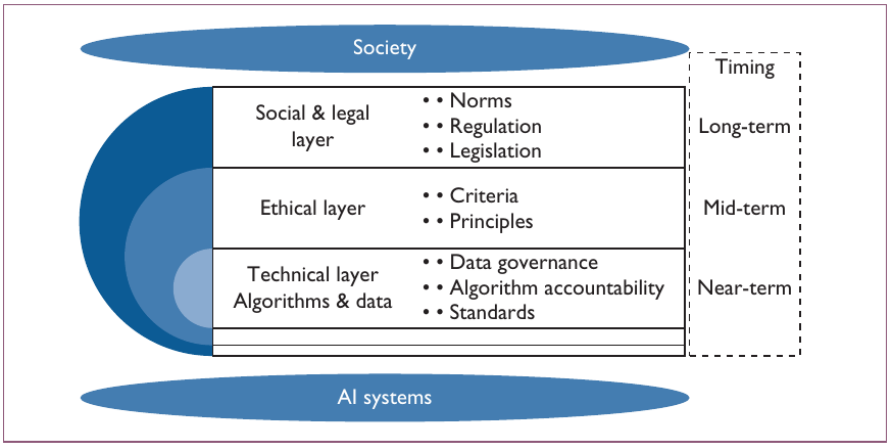

Figure 1: A layered model for AI governance. The interacting layers (which sit between society and AI applications) are social and legal; ethical; and technical foundations that support the ethical and social layers.

Figure 1 shows a representation of the layered governance model. It will sit between society and AI applications. The instruments mapped onto the layers can be developed at different times. In the near term, governance proposals could concentrate on developing standards and principles for AI algorithms. For the mid- and long-term, nation-states can work on specific legislation to regulate mature AI applications. The model can be a helpful heuristic that illustrates how principles, policies, norms, and laws in response to AI-based challenges and opportunities can be combined and work together, within and across layers.

The Technical Layer

The technical layer is the foundation of the AI governance ecosystem — the algorithms and data out of which it is built. AI systems and autonomous systems rely on data and algorithms, regardless of whether they are physical systems (such as self-driving cars and commercial robots) or software systems (such as criminal justice or medical diagnostic systems, or intelligent personal assistants)[17]. A set of principles for accountable algorithms and an associated suggested social impact statement were developed as part of a Dagstuhl Seminar on "Data, Responsibly."[18]. The proposed principles for accountable algorithms with social impact are as follows: responsibility, explainability, accuracy, auditability, and fairness. The collection, use, and management of data by AI algorithms, known as data governance, should follow principles that promote fairness and safeguard

against race, color, national origin, religion, sex, gender, sexual orientation, disability, or family status discrimination[19].

The Ethical Layer

On top of the technical layer, we could articulate high-level ethical concerns that apply to all types of AI applications and systems. One important source for the development of such ethical principles are human rights principles. Another example of the emergence of AI ethics norms is the IEEE general principles for AI and autonomous systems[17]. Actions driven by algorithms can be assessed according to ethical criteria and principles. For instance, when an AI application analyzes the data of an insurance company and charges a certain group of people higher premiums, based on variables such as gender or age, such a decision-making application would be violating the ethical principle of equal or fair treatment.

The Social and Legal Layer

The social and legal layer could address the process of creating institutions and allocating responsibilities for regulating AI and autonomous systems. For example, Matthew Scherer[20] describes a policymaking body that would have the power to define AI, create exceptions allowing for AI research to be conducted in certain environments without the researchers being subjected to strict liability, and establish an AI certification process. One starting point for specific norms aimed at regulating AI can be the principles and criteria that emerge from the ethical and technical layers, in addition to pre-existing and more general national and international legal frameworks, including human rights. The layered model provides a framework for thinking about AI governance, aiming at the definition of appropriate behavior for AI and autonomous systems.

END

Implementing governance structures for AI and algorithmic decision-making systems can occur at multiple layers and involve blended approaches. Here, we describe some of these layers, taking into consideration that some of them would only be considered if the risk that certain AI applications present are substantial and concrete. Governance processes can range from market-oriented solutions to government-based structures and can be applied nationally or internationally. On the regional level, a rich example is the General Data Protection Regulation (GDPR), a wide-ranging and complex regulation intended to strengthen and unify data protection for all individuals within the European Union (). It offers a (limited) “right to explanation” that will oblige companies to explain the purpose of an algorithm and the kind of data it uses when making automated

decisions[21]. Absent an AI-specific international legal framework, a global oversight body, which can take the form of a multistakeholder committee, could be the curator of global principles and emerging norms for AI systems.

References

1. E. Horvitz, "AI, People, and Society," Science, vol 357, no. 6346, 2017, p. 7.
2. National Science and Technology Council Committee on Technology, Preparing for the Future of Artificial Intelligence, tech. report, Executive Office of the President, 2016.
3. S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accoun able AI for Robotics," Science Robotics, vol. 2, no. 6, 2017; doi:10.1126/scirobotics.aan6080.
4. S. Levin, "New AI Can Guess Whether You're Gay or Straight from a Photograph," The Guardian, 8 Sept. 2017; .
5. I. Rahwan, "Society-in-the-Loop: Programming the Algorithmic Social Contract," Ethics and Information Technology, 2017; doi:10.1007/s10676-017-9430-8.
6. P. Stone et al., "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence," Report of the 2015 Study Panel, tech report, Sept. 2016; .
7. J. McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 31 Aug. 1955.
8. E. Kumar, Artificial Intelligence, I.K. International, 2008.
9. R. Brownsword and K. Young, eds., Regulating Technologies: Legal Futures, Regula- tory Frames, and Technological Fixes, Hart, 2008.
10. R. Calo, "Artificial Intelligence Policy: A Roadmap," Social Science Research Network (SSRN), 8 Aug. 2017; .
11. J. Palfrey and U. Gasser, Interop: The Promise and Perils of Highly Interconected Systems, Basic Books, 2012.
12. C. Scott, "Regulation in the Age of Governance: The Rise of the Post-Regulatory State," The Politics of Regulation: Institutions and Regulatory Reforms for the Age of Governance, J. Jordana and D. Levi-Faur, eds., Edward Elgar, 2004, pp. 145–174.
13. R.H. Weber, Realizing a New Global Cyberspace Framework: Normative Foundations and Guiding Principles, Schulthess 2014.
14. U. Gasser, "AI and the Law: Setting the Stage," Medium, 26 June 2017; .
15. C.S. Yoo, "Protocol Layering and Internet Policy," Faculty Scholarship Paper 454, Univ. of Pennsylvania, 2013; .
16. D. Clark, "Characterizing Cyberspace: Past, Present, and Future," MIT CSAIL, v. 1.2, 12 Mar. 2010.
17. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous

Systems (AI/AS), IEEE, 2017; .
18. S. Abiteboul et al., "Data, Responsibly (Dagstuhl Seminar 16291)," Dagstuhl Reports, vol. 6, no. 7, 2016, pp 42–71.
19. National Science and Technology Council Committee on Technology, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, tech. report, Executive Office of the President, 2016.
20. M. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," Harvard J. Law & Technology, vol. 29, no. 2, 2016; .
21. S. Wachter, B. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," Int'l Data Privacy Law, 2017; .

- **Urs Gasser** is the executive director of the Berkman Klein Center for Internet & Society at Harvard University, where he is co-leads the Ethics and Governance of AI initiative, and serves as a professor of practice at Harvard Law School. His research and teaching focus on the interplay between law and technology. Gasser is a graduate of the University of St. Gallen and Harvard Law School. Contact him at .
- **Virgilio A.F. Almeida** is a faculty associate at the Berkman Klein Center for Internet and Society at Harvard University, and a professor in the Computer Science Department at the Federal University of Minas Gerais (UFMG), Brazil. His research interests include cyber policies, large-scale distributed systems, the Internet, and social computing. Almeida has a PhD in computer science from Vanderbilt University. Contact him at or .