BROOKINGS

COMMENTARY

The AI regulatory toolbox: How governments can discover algorithmic harms

Alex Engler

October 9, 2023

→ While AI legislation advances, some regulators are experimenting with gathering information about algorithmic systems and their potential societal effects.

→ This experimentation has developed a toolbox of AI regulatory strategies, each with different strengths and weaknesses.

→ These potential interventions include transparency requirements, algorithmic audits, AI sandboxes, leveraging the AI assurance industry, and welcoming whistleblowers.

Governments around the world are implementing foundational policies to regulate artificial intelligence (AI) and algorithmic systems more generally. While legislation is advancing, regulators should not wait idly for legislators to act. Instead, regulators should be actively learning about the algorithmic systems in their regulatory domain and evaluating those systems for compliance under existing statutory authority.

Many regulatory agencies have started this work, including the U.S. Federal Trade Commission's (FTC) Office of Technology ↗ and Consumer Financial Protection Bureau (CFPB), new algorithmic regulators in the Netherlands ↗ and Spain ↗, and online platform regulators such as the UK's Office of Communications ↗ (OFCOM) and the European Centre for Algorithmic Transparency ↗. These agencies and others have started to implement novel approaches and policies for AI regulation.

Of particular interest is how oversight agencies can learn about algorithmic systems, as well as their societal impact, harms, and legal compliance. As agencies experiment in gathering this information, it is possible to broadly characterize an emerging AI regulatory toolbox for evaluating algorithmic systems, particularly those with greater risk of harm.

The toolbox includes expanding transparency, performing algorithmic audits, developing AI sandboxes, leveraging the AI assurance industry, and learning from whistleblowers. These interventions have different strengths and weaknesses for governing different types of AI systems, and further, they require different internal expertise and statutory authorities. To better inform AI policymaking, regulators should be aware of these tools and their trade-offs.

1. *Expand Algorithmic Transparency Requirements*

Mandating corporate disclosures is a key function of many government agencies, and this role is also valuable in markets of algorithmic systems. Algorithmic transparency is among the most thoroughly studied subfields of AI, which has resulted in a wide variety of approaches, including transparency measures for affected individuals, the general public, and to other organizations, such as other businesses or regulators themselves.

In its simplest form, transparency for affected individuals means direct disclosure, in which an individual is informed when they are interacting with an algorithmic system. However, it also includes "explainability," in which an individual is offered some insight into why an algorithmic system generated a specific decision or outcome. Public-facing transparency might include statistics about the outcomes of an algorithmic system (such as how accurate or how fair it is), descriptions about the underlying data and technical architecture of an algorithmic system, and a more comprehensive examination of its impacts, often called an algorithmic impact assessment ↗. Lastly, transparency can be cross-organizational, such as between two businesses or between businesses and regulators. Because this is not fully public, it may enable more detailed information sharing, such as proprietary information about algorithmic systems, which may be helpful for the clients of AI developers who seek to adapt an algorithmic system ↗ for a new purpose. Similarly, regulators may be able to seek more specific information privately, reducing risks to intellectual property theft while enabling the government agency to better understand a type of algorithmic system.

Regulators may also have existing authorities that can be applied to mandate or encourage algorithmic transparency. The CFPB has stated clearly that legally required explanations of credit denial apply to algorithmic systems ↗. Also in the U.S., longstanding interagency guidance ↗ and a new proposed rulemaking ↗ require financial lenders to assess information about any algorithmic systems for property valuation that they procure. The European Union's (EU) General Data Protection Regulation (GDPR) guarantees ↗ an individual right to "meaningful information about the logic" of algorithmic systems. This has led companies, such as home insurance providers ↗, to offer responses—albeit limited ones—to requests for information about algorithmic decisions. Although not yet passed into law, the forthcoming EU AI Act will also create substantial new transparency requirements (https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/#anchor5) , likely including direct disclosure of chatbots and public reporting about high-risk AI systems.

Transparency requirements require little expertise and capacity from government agencies, making them an appealing early step in AI regulation. However, regulators do need to be careful in specifying transparency requirements—vaguely or poorly worded requirements can permit too much flexibility in algorithmic transparency, allowing for companies to cherry pick self-serving statistics.

When transparency requirements are sufficiently narrowly tailored to a type of algorithmic system, they can lead to a wide range of benefits. Public information about an algorithmic function can help individuals and other businesses make better choices about which AI developers to patronize or work with. AI developers may themselves realize from public disclosures that their systems are not performing at the state of the art, leading them to prioritize product improvements. Yet, even if more transparency does not lead to introspection, public information can help journalists and civil society organizations identify subpar and potentially harmful systems. Journalism can lead to public scrutiny that leads to change in the practices of AI developers, while civil society organizations may make use of lawsuits to punish lax algorithmic practices. All these benefits can arise even without the regulators themselves using public information, although transparency also helps inform better policymaking and other interventions from the AI regulatory toolbox.

2. *Performing Algorithmic Investigations and Audits*

An especially impactful approach for regulators is performing algorithmic investigations and audits (hereafter audits, for simplicity), which are evaluations of an algorithmic system. There is a growing body of scientific research ↗ on how algorithmic audits can be conducted and what they can discover. Audits have revealed inaccuracy, discrimination, distortion of information environments, misuse of underlying data, and other significant flaws in algorithmic systems. Depending on the country and the algorithmic application, these flaws can be in violation of laws and regulations that might be of interest to regulators.

Algorithmic audits may become a common tool of AI regulation. The core enforcement mechanism of the EU's AI Act is the ability of regulators to demand information on high-risk algorithmic systems to assess compliance with the law. Beyond the many algorithmic audits from academics ↗ and journalists ↗, some regulators have already begun to use this oversight tool. The Australian Competition and Consumer Commission performed an audit ↗ of Trivago's hotel ranking advertisement and found it was misleading consumers. Both the United Kingdom's (UK) Information Commissioner's Office (ICO) and the Competition and Markets Authority have also engaged in algorithmic audits ↗ aimed at algorithms on, or using data from, online platforms. Further, the UK's OFCOM is actively staffing up in order to enable future algorithmic audits as part of the Online Safety Bill. In the U.S., the FTC has the legal authority to audit algorithmic systems through its information gathering tools, such as the civil investigative demand.

Audits are particularly well suited to discovering algorithmic flaws because they do not require trusting the claims of an algorithmic developer, but rather enable direct analysis by regulators. Notably, algorithmic audits can range in how involved they are ↗, from a limited audit in which auditors only review documentation, all the way to a comprehensive inspection of the specific technical function, outputs, and broader sociotechnical deployment of an algorithmic system. These more intensive algorithmic audits are far more likely (https://www.brookings.edu/articles/auditing-employment-algorithms-for-discrimination/) to uncover flawed and harmful aspects of an algorithmic system. However, more intensive algorithmic audits are also far more technically complex, requiring more expertise and technical capacity from regulators. Specifically, regulators would need data scientists with expertise in evaluating algorithmic systems and may need to take necessary steps to develop a computing environment for algorithmic evaluation with appropriate privacy and cybersecurity safeguards.

3. *Develop Regulatory AI Sandboxes*

An AI regulatory sandbox is meant to systematically improve communication between regulators and regulated entities, most frequently AI developers. Participation in AI sandboxes, which is often voluntary, is meant to ease regulatory compliance and offer legal certainty to companies while improving regulators' understanding of the design, development, and deployment of a type of AI system. This may also help regulators identify potential legal problems with a particular AI system during its development. In addition to preventing harms, this can enable an AI developer to make earlier—thereby potentially less costly—course corrections on its algorithms.

There is no specific technical definition of an AI sandbox; the term can refer to a range of approaches from a simple ongoing exchange of documentation (from companies) and feedback (from regulators) all the way to a computing environment shared by a company and regulators. This creates some uncertainty—for instance, while the European Parliament's version of the AI Act requires each EU member state to establish at least one regulatory sandbox, it is not clear what precisely each country would implement.

The first such AI sandbox has been recently launched ↗ by a partnership between the European Commission and the Spanish government, but regulatory sandboxes for other industries are not new. Over 50 countries have experimented with using regulatory sandboxes for digital financial services ↗, and the OECD has documented ↗ others in biotechnology, health, energy, and waste treatment. Some of these sandboxes have performed assessments on AI systems, such as those from the UK ICO and the U.S. Consumer Financial Protection Bureau ↗. Particularly valuable are the public reports published by the ICO when a financial technology application leaves the regulatory sandbox. These reports can include detailed information ↗ on how a specific algorithmic system can comply with regulatory requirements, thereby informing the public and other companies building similar applications.

AI sandboxes have many distinguishing qualities relative to other AI regulatory interventions. First, they require ongoing collaboration between regulators and regulated companies, and may be less adversarial than an algorithmic audit. Sandboxes may require more work for companies (such as sharing updated data or ensuring an algorithmic system works in a government computing environment).

However, they also provide more clarity to the AI developers, who may receive feedback earlier and more frequently on regulatory compliance. In some cases, this could accelerate time-to-market, especially under a governance regime with ex-ante or pre-market requirements.

AI regulatory sandboxes can also demand more from regulators, especially if they entail developing a computing environment. Beyond the skills necessary for algorithmic auditing, regulators would need to ensure that their computing environments can accommodate a broad range of algorithmic software in order to allow various AI developers to use the sandboxes. Further, regulators may have to develop regulatory sandboxes that are capable of testing many distinct types of algorithmic systems, including algorithms built into phone apps, online platforms, and physical products. Holding algorithmic systems indefinitely in government computing environments during development may increase risks to intellectual property, increasing the stakes of strong cybersecurity. Due to the significant workload required for AI sandboxes, they may be more appropriate for relatively high-stakes algorithmic systems.

4. *Leverage the AI Assurance Industry*

AI assurance is a catchall term for a variety of technology companies that specialize in monitoring, evaluation, and legal compliance of algorithmic systems. There are many companies in this emerging market, including Weights & Biases, Babl AI, Eticas Research and Consulting, Credo AI, Fairly AI, SolasAI, Fiddler AI, FairPlay AI, Armilla AI, Trustible, and Holistic AI. Although their offerings may overlap, their business models vary significantly. Some companies, such as Weights & Biases, offer bespoke software that primarily aids in the algorithmic development process. However, these tools also enable documentation and storage of past data and models, which leads to the reproducibility that is necessary ↗ for detailed regulatory compliance. Other companies, such as Trustible, are primarily focused on documenting algorithmic systems and their compliance with specific standards or regulations, without offering developer tools. Some are industry specific—Fairplay AI focuses narrowly on fairness and disparate impact analyses for financial institutions. Others, such as Eticas Consulting and Babl AI, offer full algorithmic audits and associated compliance services, aiming to improve fairness but also performance and safety more generally.

A common strand across the entire AI assurance industry is a mixed business model that advertises both profit-motivated improvements to algorithmic systems and better preparedness for regulatory and legal compliance. For instance, several AI assurance companies stress the value of internal monitoring, so corporate leaders can understand and scrutinize the function of their own algorithms, in addition to highlighting future legal requirements. This likely a stronger sales pitch to potential clients, especially given that most AI laws are still being drafted, rather than being implemented.

Although this industry is distinct from governance, regulators should actively engage with the AI assurance industry to advance democratic goals, perhaps best exemplified by the U.K ↗. Regulators can issue guidance that encourages regulated companies to consider using AI assurance tools, even possibly noting this could be interpreted as a potential signal of regulatory compliance. Further, regulators can inform and learn from the AI assurance industry. By communicating about specific technical functions and the societal impacts of algorithmic systems in a regulated field, regulators can help AI assurance companies strive towards not just nominal compliance, but meaningfully better outcomes. For instance, regulators concerned with discrimination could encourage relevant AI assurance companies to offer alternative candidate algorithms ↗ that might be less discriminatory instead of simply detecting biased results. Further, regulators can encourage and highlight AI assurance companies that establish processes which enable some degree of independent scrutiny, such as with consistent evaluation standards, although this is challenging to do when AI assurance companies depend on AI developers for revenue.

5. *Welcome Complaints and Whistleblowers*

Regulators should also welcome information from affected individuals and whistleblowers from AI developers—both of whom may have unique information about algorithmic systems.

Individuals who are subjected to algorithmic systems may have specific insight into the function of those systems. Several U.S. agencies, such as the Equal Employment Opportunity Commission, explicitly welcome reporting ↗ of discrimination from AI systems and can use those complaints to start formal investigations. However, one notable shortcoming of individual complaints is that it is often difficult or impossible for an individual to meaningfully recognize that an action by an algorithmic system was wrong or unfair for them. The infamous obscurity of algorithmic systems can make this very hard for individuals. However, groups of people have come together to identify algorithmic harms. For example, a group of content creators documented ↗ that YouTube appeared to be demonetizing their videos when the titles included LGBTQ-related vocabulary. While this is not guaranteed to be included in the final version of the law, one version of the EU AI Act includes a path to redress (https://www.brookings.edu/articles/key-enforcement-issues-of-the-ai-act-should-lead-eu-trilogue-debate/) for people harmed by algorithmic systems. Agencies should welcome these types of complaints and concerns from affected persons.

There is one group of individuals who are likely to have an intimate and sophisticated understanding of algorithmic systems—the developers themselves. Often, the data scientists and machine-learning engineers who build algorithmic systems are by far the best placed to understand their societal impact, harms, and even legal violations. Most famously, Frances Haugen provided regulators and journalists with thousands of pages ↗ of Facebook's internal documents that contradicted the company's public statements. Peter Zatko's complaints, including that Twitter enabled far too many employees access to sensitive user data, led to congressional hearings ↗ and increased scrutiny, just as Haugen's did.

While these examples are specific to online platforms, in other fields, such as financial oversight, regulators even offer cash rewards for whistleblowers ↗. Regulators should recognize when their information-gathering approaches may be systemically limited from the outside and consider the role of direct reporting and whistleblowers for discovering algorithmic harms.

Agencies should use the tools they have to understand and regulate AI

Regulators should actively consider what steps are necessary and valuable in their domains to ensure their regulatory mission is preserved. This includes cataloging and observing emerging uses of algorithmic systems in their field, exploring what their existing statutory authority allows for, and hiring staff with expertise in algorithmic systems. Regulators may benefit from a gap analysis—identifying where current authorities and capacities are lacking so that they can inform legislators, who are far less likely to understand the nuances of every regulatory subfield.

While regulators may often lack the most appropriate and best suited tools for information gathering about algorithmic systems, many will have some authority to perform information gathering. Beyond the interventions explored here, regulators can also learn from independent academic research ↗, which is especially helpful to understand algorithms as part of large online platforms ↗. Some governments, including the EU through the Digital Services Act, are even requiring access to platform data for independent researchers—this research is expected to inform regulatory investigations and even enforcement actions (https://www.brookings.edu/articles/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/) . In fact, regulators may turn to existing academic research first, even to prioritize what other information gathering tools—like those discussed here—to employ.

While algorithmic systems have become widely used in many regulated markets, these algorithms are unique to their circumstances (https://www.brookings.edu/articles/a-comprehensive-and-distributed-approach-to-ai-regulation/) . As a result, regulators need to build robust and persistent strategies to gather information for informed policymaking, oversight, and enforcement actions. Collectively, the emerging efforts of these agencies will continue to compose a regulatory toolkit upon which much future AI governance will be built.

AUTHORS

Alex Engler Fellow - Governance Studies, Center for Technology Innovation

Acknowledgements and disclosures

Google and Meta are general, unrestricted donors to the Brookings Institution. The findings, interpretations, and conclusions posted in this piece are solely those of the author and are not influenced by any donation.