Lévesque, Maroussia

**Working Paper**

# Scoping AI governance: A smarter tool kit for beneficial applications

CIGI Papers, No. 260

**Provided in Cooperation with:**
Centre for International Governance Innovation (CIGI), Waterloo, Ontario

*Suggested Citation:* Lévesque, Maroussia (2021) : Scoping AI governance: A smarter tool kit for beneficial applications, CIGI Papers, No. 260, Centre for International Governance Innovation (CIGI), Waterloo, ON, Canada

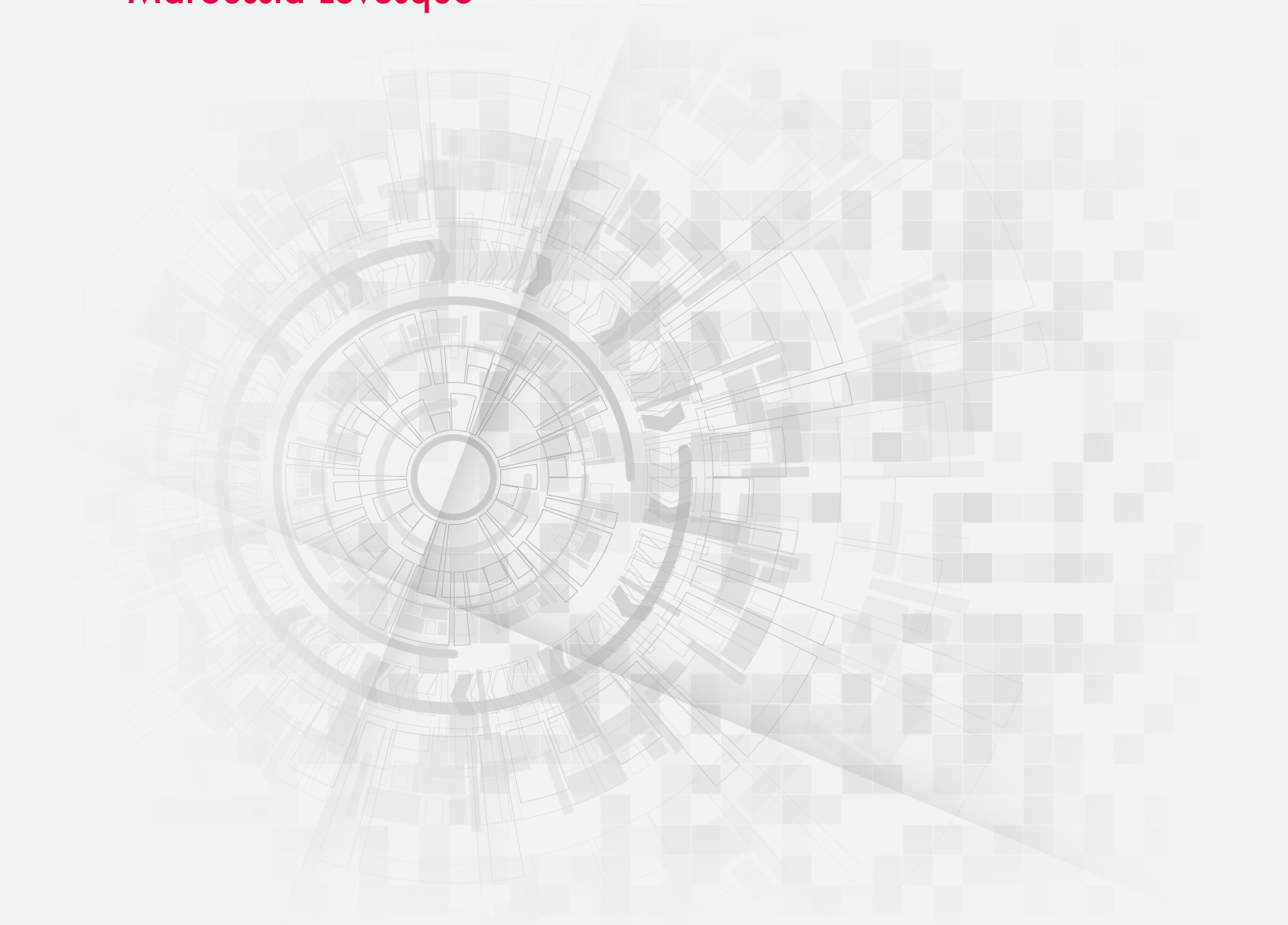
This Version is available at:
https://hdl.handle.net/10419/299732

Centre for International
Governance Innovation

CIGI Papers No. 260 — December 2021

# Scoping AI Governance
## A Smarter Tool Kit for Beneficial Applications

Maroussia Lévesque



Centre for International
Governance Innovation

CIGI Papers No. 260 — December 2021

# Scoping AI Governance
## A Smarter Tool Kit for Beneficial Applications

Maroussia Lévesque

## About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

## À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

## Credits

Managing Director of Digital Economy Robert Fay
Program Manager Aya Al Kabarity
Publications Editor Susan Bubak
Senior Publications Editor Jennifer Goyder
Graphic Designer Brooklynn Schwartz



For publications enquiries, please contact publications@cigionline.org.




Printed in Canada on Forest Stewardship Council® certified paper containing 100% post-consumer fibre.

Centre for International Governance Innovation and CIGI are registered trademarks.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

# Table of Contents

# About the Author

Maroussia Lévesque is a CIGI senior fellow, a doctoral candidate at Harvard Law School and an affiliate at the Berkman Klein Center on Internet and Society. She researches artificial intelligence (AI) governance.

Maroussia contributes to the Institute of Electrical and Electronics Engineers standard on algorithmic bias and the Indigenous Protocol and Artificial Intelligence Working Group. She previously led the AI and human rights file at Global Affairs Canada and consulted for the Global Partnership on AI. She holds degrees from Concordia University, McGill University and Harvard University, is a member of the Quebec Bar and clerked for the Chief Justice at the Quebec Court of Appeal.

# Executive Summary

Policy makers can intervene directly and indirectly to shape beneficial artificial intelligence (AI) systems. Direct action consists of imposing specific fairness constraints to redress harmful biases. However, mutually exclusive definitions of fairness hamper the viability of a top-down approach. In a regulatory landscape fraught with wide contextual variability, imposing a monolithic technical implementation of fairness could have unintended consequences. That decision is best left to actors implementing AI systems on the ground, as they can better grasp context-sensitive factors. That said, policy makers should (and, indeed, must) retain oversight of AI systems as they would for any private activity. Indirect approaches that promote procedural safeguards are most amenable to success because they foster accountability and integrity without misdirecting coarse intervention in the complex and rapidly evolving AI space. Those systemic interventions cluster around two themes: ramping up intervention with traditional regulatory tools and reinventing the role of public actors in light of industry's traction.

Ramping up regulatory intervention involves:

- imposing a presumption of discrimination for opaque systems;
- drawing red lines against abusive applications;
- attracting private talent to staff administrative agencies with specialized expertise;
- devising tax incentives to promote progressive fairness metrics;
- stimulating public infrastructure and research and development (R&D) to counterbalance commercial interests;
- compelling disclosures to support public debate; and
- setting up a no-fault compensation regime to redistribute the cost of harms more equitably.

Public actors can also organize power among private actors to create a system of checks and balances. Examples include the European Union's proposed Artificial Intelligence Act (AI Act) creating an ecosystem among producers and conformity assessors, and technical standards fostering productive collaboration between private actors.

# Introduction

Like clean air and safe roads, fair AI systems contribute to the public good. They bolster trust in innovation, alleviate a sense of exclusion from those historically marginalized and ultimately contribute to social peace. But AI's centre of gravity rests with industry. Its commercial incentives do not necessarily align with broader societal interests. In theory, public actors are uniquely positioned to promote such interests, including equality and non-discrimination. But in practice, they often sit in the back seat when it comes to designing and implementing AI policies. This paper seeks to narrow the gap between theory and practice by providing policy makers with the tools to actively shape AI governance. A jurisdiction-agnostic definition of policy makers includes members of government, legislators and civil servants supporting legislative and executive decision making.[1]

Policy makers have two options to intervene in AI governance: using traditional regulatory tools to constrain private entities and adapting to the reality of industry-driven AI development. This paper details both avenues, striving to spark a discussion among policy makers as to what approach best suits the specific issue they face, while opening the aperture regarding the array of tools available to foster beneficial AI systems. An overview of fairness harms anchors the discussion in the section titled "Equality Challenges in AI Systems." The next section, "Redress," explores light-touch oversight through classic regulatory tools such as certification schemes, tax credits and mandatory disclosures. The section titled "Adapt" invites policy makers to rethink their role vis-à-vis private actors, with the

1 While courts arguably engage in policy making as well, the oblique nature of their influence warrants exclusion from the scope of this paper. See Howard and Steigerwalt (2012).

European Union's AI Act and technical standards illustrating co-regulation. The conclusion explores combining these approaches for optimal impact.

## Equality Challenges in AI Systems

AI and machine-learning systems assist or make decisions in a plethora of contexts ranging from the mundane to the life changing. Applications include targeted advertising (National Fair Housing Alliance 2021), predicting tenant reliability, welfare fraud (Eubanks 2018) and criminal recidivism predictions (Angwin et al. 2016).[2] In all these instances, researchers, activists and journalists have documented disparate impact on racialized, low-income or otherwise minoritized groups.

In the past five years, a new field of research has emerged to address harmful biases in AI systems[3] (Altman, Wood and Vayena 2018).[4] Fairness constraints on data inputs or on results can alleviate disparate impact (Celis, Keswani and Vishnoi 2020; Kleinberg, Mullainathan and Raghavan 2016; MacCarthy 2016).[5] Many such approaches intervene on the distribution of errors among groups, attempting to equalize the performance of AI systems between majority and minority groups (MacCarthy 2016). The red cells in Table 1 represent two types of errors that result from misclassification. Type I errors consist of false positives, that is, predictions wrongly indicating a match between the case at hand and the sought-after characteristic. For example, a computer vision system that misclassifies a blueberry muffin as a Chihuahua is a type I error.[6] Conversely, type II errors entail false negatives, with results wrongly suggesting a negative result. A loan assessment algorithm that rejects a qualified applicant returns a type II error. Put another way, type I errors lead to overinclusive results, with more predictions qualifying than in reality, while type II errors generate underinclusive outcomes, with fewer instances qualifying.

Table 1: Error-Type Matrix

| Predicted/ Actual Value | Positive | Negative |
|---|---|---|
| Positive | True positive | False positive (type I error) |
| Negative | False negative (type II error) | True negative |

*Source:* Author.

Depending on the nature of the prediction, both overestimation and underestimation can be deleterious. The Chihuahua example is a fairly benign false positive, but overestimation in other applications, such as probabilistic DNA matching, can lead to wrongful convictions (Shaer 2016). Conversely, type II false negatives underestimating creditworthiness for bank loan applications can deny qualifying borrowers life-changing opportunities.[7] These errors not only harm individuals but also have implications at the population level. Indeed, errors that disproportionately affect certain groups can perpetuate vicious cycles of exclusion and systemic discrimination.

To further complicate matters, different approaches to redressing imbalances across groups are mutually exclusive. To (over)simplify, maintaining the predictive value across groups is incompatible with equalizing error rates. This is because groups typically have different base rates of the sought-after characteristic. Recidivism prediction algorithms illustrate the issue, as debates about the detrimental impact on Black offenders reveal the trade-off between fairness metrics. Before delving into the matter further, a note about the mechanics of recidivism scores is in order. The description of error types so far assumed a categorical binary positive/negative classification, but recidivism algorithms issue risk scores on a scale of one to 10. Yet the problem remains the same: whether predicting a binary classification or a score on a spectrum, algorithms can over- or underestimate outcomes. Returning to recidivism

2 *Department of Housing and Urban Development v Facebook, Inc,* FHEO No 01-18-0323-8 (2019), online: <www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf> [*HUD*].

3 See https://perma.cc/G4TK-W734 and https://perma.cc/3BMK-HQBF.

4 For a critique of narrow technical approaches, see Green (2018).

5 See https://pair-code.github.io/what-if-tool/.

6 See Cave (2016).

7 Ibid.

scores, external reviewers suggested that a widely used algorithm systematically overestimated the recidivism risk of Black offenders while underestimating that of their white counterparts.[8] That critique took issue with the unequal rate of false positives and false negatives along racial lines.

Proponents of the system retorted that Black offenders have higher rearrest rates (Kleinberg, Mullainathan and Raghavan 2016; Corbett-Davies et al. 2016; Koepke and Robinson 2017, note 127), such that the predictions of reoffence for that group will inevitably be fuzzier and thus involve more false positives. In other words, the only way to ensure the algorithm does not overestimate the recidivism risk of Black offenders would be to artificially lower their score. In their view, the predictions were fair because the system maintained predictive value across different groups, with scores conveying the same probability of reoffence irrespective of race. The debate about the fairness of recidivism prediction captures the trade-off between maintaining predictive value and equalizing error rates across groups with different characteristics. The differing baseline between groups, in this case rearrest rate, is at best a warped proxy for the "ground truth" of actual reoffence. Rearrest reflects the role of race and ethnicity in policing practices (Pierson et al.), thus adding another layer of complexity to the debate.

Implementing stratospheric-level notions of fairness on the ground involves trade-offs. Policy makers aiming to improve AI fairness through direct intervention should engage more actively with thorny editorial decisions that ultimately determine the kind and intensity of fairness AI systems promote. The idea is not to champion a specific metric across the board but to inform interventions with a lucid perspective on potential and drawbacks. That said, policy makers have lighter-touch options that promote fairness at a systemic level without directly interfering with the distribution of errors or the predictive value of results across groups. The balance of this paper explores such indirect options.

8 See Angwin et al. (2016), but see Rudin, Wang and Coker (2019) for a nuanced view on race as the determinant factor, suggesting age also played a role.

# Redress

Policy makers can ramp up oversight of the AI industry to ensure that innovations comport with the well-being of people on the ground. This section explores various avenues to do so.

## Rights and Liabilities

The prospect of liability can dissuade harmful activity. Regulators can allocate individuals protection to deter harmful corporate behaviour. For example, strengthening procedural protection can afford affected individuals a way to vindicate their right to equality and non-discrimination. Refusal to provide a legible algorithm could trigger a rebuttable presumption of discrimination. By way of analogy, the European Commission suggests that a respondent's failure to disclose documents or produce witnesses familiar with its conduct can lead to an inference of *prima facie* discrimination (European Commission Directorate-General for Justice and Consumers, Directorate D – Equality 2015; Wachter, Mittelstadt and Russell 2021).

This approach would address the information asymmetry between industry defendants and affected claimants. This asymmetry arises in two contexts, often conflated under the label "black box." Proprietary systems protected by trade secrets make up the first type of black box. Designed to prevent competitors from misappropriating valuable intellectual property, the evidentiary privilege for trade secrets shields witnesses from disclosing sensitive information about their technology in judicial proceedings. Manufacturers of AI systems could invoke trade secret protection to avoid disclosing granular information about the methodology of their proprietary system. Because access to this information is crucial for claimants to articulate equality issues, invoking trade secrets is likely fatal to discrimination claims. The second type of black box refers to models so complex that their process defies human understanding. Deep-learning architectures make inferences through a complex network of layered units that each makes a small calculation. With some models entailing millions of calculations, a straightforward causal explanation is beyond reach. Explainable models strive to provide *post facto* insight, answering the "why" question as to a model's behaviour in a subsequent step (Gilpin et al. 2018). While ongoing research seeks to provide insight into how a

particular decision is made in a complex AI system, this is still a nascent field fraught with uncertainty (Joshi, Agarwal and Lakkaraju 2021; Arrieta et al. 2020; Guidotti 2019). Both types of black boxes still impede a claimant's ability to articulate harm in a court-compatible narrative, denying them the opportunity to present probative evidence of harm. A rebuttable presumption of discrimination for opaque systems would disrupt the advantage of the manufacturer of AI systems, insofar as the latter can weaponize the opacity of AI systems to effectively shield itself from discrimination claims.

For all its potential, steering AI development through a rights and liabilities approach puts the onus on affected individuals to contest what is often intractable and diffuse harm. However, targeted legislative intervention can bolster the viability of this option. For example, lowering the bar for certifying class actions or funding public litigation would secure the ability to effectively vindicate individual rights, thus credibly deterring AI companies from engaging in harmful behaviour. Similarly, enhanced protection for whistleblowers can incentivize actors who are privy to privileged information to protect the rights of affected people (Katyal 2019; Kusisto and Sun 2021).[9]

## Command and Control

Legislators can impose penalties to draw red lines against harmful AI systems. By way of analogy, failure to take early regulatory steps to shape the internet effectively delegated power to the private sector (Black and Murray 2019). Similarly, the current laissez-faire approach in the AI space lets companies embed policy through the technical infrastructure of AI systems (Solow-Niederman 2019; Lessig 2006, 342–43). A robust command-and-control approach would reverse the tide, prohibiting certain practices and establishing safeguards to avoid AI-driven harm.

Personal data protection regimes can chip at exploitative data harvesting and discriminatory profiling. Regulating the data layer of the AI stack[10] could ensure system-wide protection and bypass definitional issues with AI systems (see Figure 1 for a summary of the levels of functionalities underpinning AI systems). A case in point is the European Union's General Data Protection Regulation (GDPR),[11] which sets out ground rules for collecting, sharing and processing personal data. Fines of up to four percent of a company's global revenue ensure compliance. The regulation sets out a special procedure for processing sensitive data[12] and safeguards specific to automated decision making. These safeguards include a right to information at the time of data collection and processing as well as enhanced protection for processing with legal or otherwise significant effects (Sartor and Lagioia 2020).[13]

However, regulating personal data is both underinclusive and overinclusive. It is underinclusive because some AI-driven harms materialize without personal data involved. Indeed, machine learning can make powerful inferences from trivial data points. For example, emotional recognition provides intimate insights into people's emotional state even if it does not identify them. Not only is this deleterious to personal autonomy (Cohen 2019), it may also entail discriminatory effects if systems perform differently across various ethnic racial features and cultural expressions of emotions.[14] Recent computer science advances also enable machine learning to do more with less data (Sucholutsky and Schonlau 2020; Hao 2020; Hoefler et al. 2021), further eroding the ability of data governance to comprehensively tackle AI-driven harms. Conversely, personal data regulation can be overinclusive. Limitations on collecting or processing sensitive data can prevent tracking disparate impact on protected

9 In the United States, the Securities and Exchange Commission can award 10 to 30 percent of fines over US$1 million to whistleblowers who help it investigate fraud. See https://perma.cc/KHA5-T38J. For a discussion of potential liability under non-disclosure agreements, see Kusisto and Sun (2021).

10 See Brown and Marsden (2013) for a description of the different layers of internet infrastructure; see Figure 1 for a summary of the internet and AI stack.

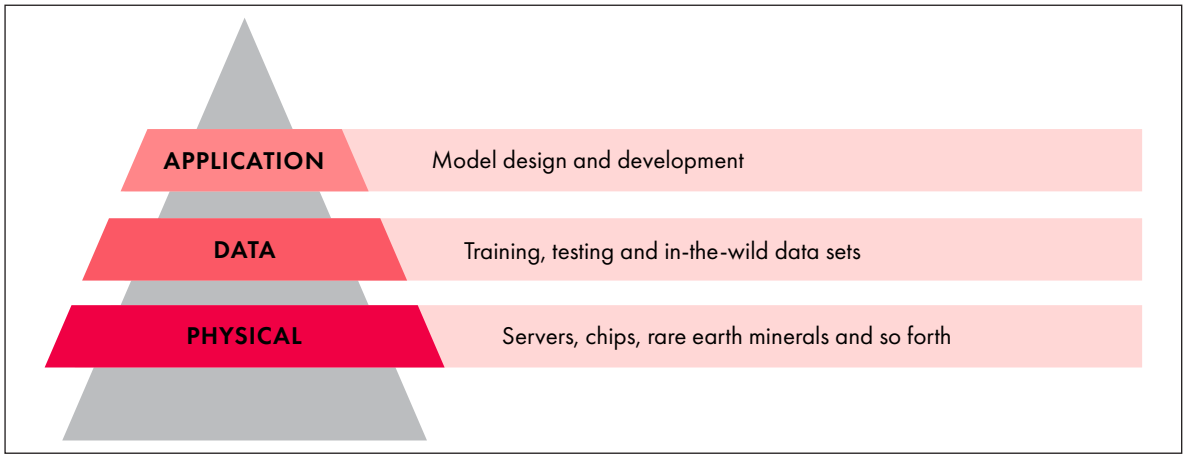
11 EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, [2016] OJ, L 119, art 9, online: <https://perma.cc/VA3J-GHV4> [*GDPR*].

12 Ibid.

13 Ibid., arts 13(2)(f), 15(1)(h), 22; for the disputed scope of the right to an explanation, see Goodman and Flaxman (2017); Wachter, Mittelstadt and Floridi (2017); Kaminski (2019a).

14 See related research by Buolamwini and Gebru (2018) regarding facial recognition underperforming on women of colour; see also McStay and Urquhart (2019), calling into question the science behind emotional sorting based on physiological markers.

## Figure 1: AI Stack




*Source:* Author.

or otherwise vulnerable groups. Absent careful calibration, overbroad personal data regulation can thus undercut efforts to identify disparate impact. In sum, data regulation is at best insufficient, at worst counterproductive when it comes to addressing equality and related concerns in AI systems. AI governance therefore requires targeted intervention at the application layer of the stack.

The European Union's proposed AI Act[15] exemplifies a targeted approach. The act delineates which AI systems are acceptable and which ones are off limits. In addition to banning subliminal, exploitative and social-scoring applications, the act restricts law enforcement's use of real-time, remote biometric identification in public spaces.[16] Given that the European Union already addresses personal data through regulations and directives, this AI-specific intervention is best understood as an add-on rather than a stand-alone regulatory strategy. Indeed, the AI Act is in dialogue with personal data regulation. For example, article 10(5) sets out an explicit carve-out from the GDPR's prohibition on processing sensitive data for tracking bias. Recital 41 conveys this rationale as follows: "In order to protect the right of others from the discrimination that might result from the bias in AI systems, the providers should be able to process also special categories of personal data, as a matter of substantial public interest, in order to ensure the bias monitoring, detection and correction in relation to high-risk AI systems."

This reference to "substantial public interest" aligns the AI Act with the GDPR, as section 9(2)(g) of the latter allows processing sensitive data when necessary for the public interest. The AI Act is a case in point of regulating at the application level of the AI stack to correct overinclusive data protection. More broadly, it is an example of juxtaposing regulatory instruments intervening at different layers of the stack to close loopholes — in this case, using incorporation by reference to align the data and application layers of the AI stack. Policy makers in other jurisdictions should take note and consider combining layer-specific and stack-wide approaches.

Detractors claim that command-and-control legislation is protectionism by another name that stifles innovation. In the personal data protection context, more than 1,000 US news sites geo-blocked EU-based users to avoid hefty GDPR fines, suggesting some empirical basis for that claim (Masnick 2018; South 2018). Similarly, a think tank backed by US companies projects that the AI Act will cost the European economy €31 billion in the next five years and reduce AI investments by 20 percent.[17] Former Alphabet/Google executive Eric Schmidt further criticized

15 *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts,* [2021] COM/2021/206 final, online: <https://perma.cc/H42G-AB3Q> [*AI Act*].

16 Ibid., art 5.

17 See Mueller (2021); but see Haataja and Bryson (2021), challenging the report's estimated 17 percent markup and projecting a five percent cost instead.

the act's transparency requirements,[18] warning that this regulate first, innovate later approach hinders competition with China (Haeck 2021). With many economies looking to reboot competitiveness through AI-related innovations, this powerful narrative is an uphill battle for those looking to foster rights-respecting machine learning in light of the trump card of economic development. However, framing AI development as an inevitable race pitting the West against China is a battle that rights-conscious democracies will inevitably lose. The racing mentality begs the question: race toward what? This discourse normalizes externalizing uncertainty about harms just to secure first-mover advantage. Furthermore, regulation fosters predictability, which is conducive to innovation, as companies can move along the translational timeline with confidence regarding their product's compliance. While there is room for caution against heavy-handed command-and-control regulation, a healthy dose of skepticism regarding those decrying its deleterious effects remains in order.

## Administrative Oversight

Specialized administrative agencies can foster nimble oversight of AI applications. A US Federal Drug Agency-like entity could modulate the type and intensity of oversight according to the complexity and transparency of algorithms (Tutt 2017). For instance, a pre-market approval scheme would certify mission-critical algorithms and self-driving cars. Administrative certification could also modulate the standard for civil liability, with negligence applying to certified AI systems and no-fault strict liability to uncertified ones (Scherer 2016). Certified AI applications would be held to a lower standard, with the plaintiff having to prove fault. By contrast, uncertified applications gone awry would attract liability irrespective of fault. Agencies can also base certification on a precautionary approach. Where there is uncertainty about serious and irreversible harm, the proponent of an application bears the burden of proving safety.[19] This approach would capture harms beyond discrimination, allowing regulators to holistically consider the risk of harm.

Factors modulating the intensity of administrative action cluster around uncertainty, complexity, transparency and impact of AI systems. Those factors can be brought under the umbrella of a risk-based strategy. The following factors inform the timing and type of regulatory intervention:

- compensable individual risk (*ex post* liability through a develop-deploy-regulate approach);
- high individual risk (authorization through a develop-regulate-deploy approach);
- non-compensable deep-regret individual risk (through licensing, ongoing monitoring, enforcement and consent for exposure to risk);
- compensable systemic risk (*ex ante* regulation through a develop-regulate-deploy approach and *ex post* remedy through a develop-deploy-regulate approach); and
- non-compensable systemic risk (highly restrictive *ex ante* regulation on development and deployment, as well as trialling [for example, genetically modified organisms, stem cells, aviation, nuclear power]) (Black and Murray 2019, 13).

*Ex post* enforcement is especially apposite for compensable individual risk, but also complementary to *ex ante* intervention for non-compensable deep-regret individual risk and compensable systemic risk. AI systems with a disparate impact would fall into the latter category, insofar as discrimination is a theoretically quantifiable systemic risk.

Administrative oversight is not a one-size-fits-all solution. A risk-based strategy can focus regulators on problematic AI applications and avoid heavy-handed blanket intervention that unduly stifles innovation. In fact, administrative agencies already

18 *AI Act*, *supra* note 14, art 13 ("High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.")

19 For applications in the international environmental law context, see *Rio Declaration on Environment and Development*, 3–14 June 1992, UN Doc A/CONF.151/26 (vol. 1), 31 ILM 874, Principle 15 (entered into force 12 August 1992); *Convention on Biological Diversity*, 5 June 1992, UNEP/CBD/COP/1/17, Preamble (entered into force 29 December 1993); *United Nations Framework Convention on Climate Change*, 3–14 June 1992, art 3.3 (entered into force 21 March 1994); *Cartagena Protocol on Biosafety to the Convention on Biological Diversity*, 29 January 2000 (entered into force 11 September 2003).

engage with AI systems. For example, the Malta Digital Innovation Authority certifies AI systems on a mandatory or voluntary basis depending on the sector of activity (Ellul et al. 2021). With more jurisdictions likely to follow suit, the risk-based strategy outlined above provides a road map for tailoring intervention to the specific context within which specific AI systems are embedded.

Lack of specialized expertise and uncompetitive work conditions in the public sector threaten the feasibility of public oversight. Entry- and mid-level public servants typically change postings every two years or so, hindering the development of deep area expertise. Furthermore, employment conditions in public administration pale in comparison to the perks of industry players. But creative solutions can alleviate these roadblocks. An improved loan assistance repayment plan and scholarships could be tied to completing two years of public service and incentivize recent graduates to join the public sector. Governments could also "borrow" private talent in return for tax breaks or fast-track citizenship for these crossover employees. Along similar lines, a former Facebook employee proposed a dedicated oversight body amenable to former technology company employees doing a "tour of duty" to share their unique insights into internal industry processes to inform regulation (*The Wall Street Journal* 2021). A confidentiality-preserving system would buttress the arrangement, with non-disclosure agreements and security clearances organizing information exchange. However, this approach would have to be carefully calibrated to maintain the integrity of public values and avoid subverting the process into "insourcing."

## Incentives

Regulators can deploy incentives to encourage specific behaviour. Incentives can encourage voluntary alignment between industry players and public values such as equality and non-discrimination. Tax credits can reward best practices such as voluntary certification, debiased training data sets or equal-by-design systems.

A particularly promising best practice consists of optimizing for a progressive fairness metric to ensure that industry players make fairer AI systems. Perhaps the best way to define what are progressive measures is to start by explaining what they are not. The predictive parity metric mentioned in the section titled "Equality Challenges in AI Systems" is regressive because it perpetuates oppression of historically marginalized groups. Although it ensures that a criminal recidivism score (for example, six) means the same across race, it does so at the cost of accepting higher error rates for minorities. By contrast, ameliorative action attenuates the disparate impact of algorithms on protected or otherwise marginalized groups. One such measure is demographic parity, which consists of selecting the same proportion of qualified people across groups. In a university admission hypothetical, an institution that admits 70 percent of white applicants with a 4.0 or higher GPA would admit the same proportion of Black applicants with a 4.0 or higher qualification (Wachter, Mittelstadt and Russell 2021, 56). Demographic parity is not a magic bullet, as the question of *which* qualified applicant in the majority and minority pools should be selected remains unanswered. But it offers an equal chance to qualified individuals irrespective of race.

Policy makers can also leverage tax breaks to encourage greater transparency about how AI systems work. Complex or proprietary systems' opacity hinders the ability to identify, let alone resolve, equality issues (Pasquale 2016; Wexler 2018; Rudin, Wang and Coker 2019). The ability to "look under the hood" is crucial for informed public debate about the broader societal implications of AI systems. To note, simple interpretable models often perform just as well as complex ones, with the choice coming down to cost allocation (Rudin 2019). Instead of investing analyst and computational resources upstream to come up with a simple accurate model, complex models pass uncertainty to the affected users downstream. Similarly, proprietary models limit legal exposure for manufacturers, as the evidentiary privilege for trade secrets effectively shields them from liability. Unable to understand, let alone challenge and redress errors, harmed people on the receiving end absorb the error cost of flawed, opaque systems. Viewed in this light, black-box machine-learning algorithms are a form of divestment from the public good.

Policy makers can shift the tide with fiscal incentives encouraging the deployment of interpretable, open-source models. Tax breaks could reward companies that invest in developing simple models and assume the risk of legal liability and criticism by making their product open source. Tax breaks for legible models reward the internalization of error costs. They encourage companies to expend

human and computational resources to find the right predictors and construct interpretable, legible models accordingly, which ultimately reduce the possibility of externalized harms.

Tax breaks are a good option when regulators lack a legislative basis for intervening in unregulated private activities. These incentives could reach the grey market of data brokers that enable discrimination in loans or housing, and eventually solidify into standard practices for new entrants. Structured as a percentage of taxable earnings, tax breaks could be attractive even to companies focused on the bottom line.

## Market-Harnessing Controls

Policy makers can influence competition through markets. They can stimulate public R&D to infuse AI development with a non-economic agenda. The United States has recently begun exploring this option, with the National Artificial Intelligence Research Resource Task Force issuing a request for information for an Implementation Plan for a National Artificial Intelligence Research Resource. This public resource would provide access to data, computational power, educational tools and user support. Its goal is to "democratize access to the cyber infrastructure that fuels AI research and development, enabling all of America's diverse AI researchers to fully participate in exploring innovative ideas for advancing AI, including communities, institutions, and regions that have been traditionally underserved — especially with regard to AI research and related education opportunities."[20]

Policy makers can also diversify the offer at the data layer of the AI stack, encouraging data sets controlled by publicly accountable actors to ensure data provenance (Solow-Niederman 2019, 689). For instance, data trusts could counter the "dataopolies" of the "Frightful Five" (Alphabet, Amazon, Apple, Facebook and Microsoft) and rebalance the power asymmetry between data-harvesting companies and individuals who are subject to those practices (Open Data Institute 2019; Element AI and Nesta 2019). Be it through comprehensive resources or narrower data trusts, public support for AI could inject different values in the development of AI systems, fostering hospitable conditions for more equitable applications on the ground.

OpenAI CEO Sam Altman evoked failed discussions with the public sector before turning the ground-breaking non-profit organization into a capped profit structure (now essentially a Microsoft research arm with a US$10 billion investment).[21] This missed opportunity to align vanguard research with public values should spur soul-searching in public policy circles about how to better recognize, seize and create opportunities to harness scarce AI talent toward the public good.

## Public Infrastructure

Public actors can build AI systems from the ground up to foster consistency with public values and obligations. Mounting evidence suggests that constitutional and regulatory safeguards are lost in the handoff between public mandate and private implementation (Eubanks 2018; O'Neil 2016; Larson et al. 2016). Building public AI infrastructure could reverse the tide. Concretely, a law or administrative rule could establish a presumption against outsourcing AI systems in high-stakes contexts such as criminal sentencing and social benefits determinations. Insofar as public values are engrained in the culture of public institutions, keeping these systems in-house would foster equality and procedural protections. To be sure, some public institutions have a poor track record of complying with public law obligations. But the solution is to reform them in the medium to long term, not outsource processes to private actors who have even fewer incentives to promote the public good. Such infrastructure is important to develop alternatives to private applications embedding a narrower, profit-driven set of values.

## Mandatory Disclosures

Policy makers can compel the business sector to disclose the performance metrics of their AI systems. For example, they can require social media companies to release false-positive and false-negative error rates for automated content moderation. Government-issued metrics would foster harmonized reporting across companies, affording external stakeholders a meaningful opportunity to issue constructive criticism. The

20 *Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource*, 86 Fed Reg 39081 (2021).

21 See Klein (2021): "A little known fact, we tried to get the public sector to fund us before we went to the capped profit model. There was no interest."

initiative could build on existing methodologies such as the Corporate Accountability Index's algorithmic content curation indicator.[22] Sector-specific intervention can also promote disclosure. For instance, administrative agencies overseeing housing regulation can issue guidelines on automated decision making or issue non-binding standardized disclosure forms to incentivize such information sharing.[23] When adjudicating individual complaints, they can also draw a negative inference from the respondent's failure to disclose sufficient information on how an AI system works. Irrespective of the specific regulatory path, disclosures can narrow the gap between the scale and speed of automated processes and analogue conceptions of individualized procedural guarantees. Information about system-wide performance can enable public discussion and motivate companies to do better.

Companies could invoke trade secrets to resist disclosures. However, many mechanisms can support accountability without exposing proprietary information. A performance summary could be made public, with verifying details disclosed to a limited group of vetted third-party auditors bound by confidentiality undertakings. Should AI systems be litigated, hearings can proceed in camera to safeguard proprietary information. Protective measures aptly secure much more sensitive information in national security and criminal trials; they can certainly accommodate lower-stakes trade secrets.[24]

### Public Compensation

No-fault compensatory regimes can address the inherent risks of otherwise beneficial AI applications. Some AI products may entail an irreducible portion of risk, the distribution of which is impossible to predict. A case in point is automobiles, a beneficial innovation that nevertheless entails inevitable accidents. For AI applications that similarly involve social benefit notwithstanding some risks, companies could contribute a portion of earnings to a publicly administered collective liability fund that compensates impacted individuals irrespective of negligence.[25] In order to be equitable, contributions should be based on a progressive tax system. For example, companies would contribute one percent of their first $100,000 of taxable earnings; two percent of the next bracket from $100,001 to $500,000; five percent of the $500,001 to $1-million earning bracket and so on. This system would best suit randomly distributed compensable individual risk.[26] For risks that systemically fall upon one portion of the population, especially protected classes, the political choice to tacitly endorse disparate impact may not be desirable.

A compensation fund could foster moral hazard, discouraging companies to reduce harms on the rationale that their mandatory contribution prepays for externalizing harm. Policy makers should therefore combine automatic no-fault liability with the threat of negligence, allowing affected people to top-off capped compensation from the fund with additional compensatory and punitive damages (Marchisio 2021).

## Adapt

Public entities can reimagine their role in light of industry's prominence in AI. This section first articulates why industry dominance is problematic. It then explores how public actors can pivot from catching up with companies to organizing checks and balances among them.

### The Diagnostic: Corporate-Driven AI

The development and policy capacity of private AI actors dwarfs the public sector by many orders of magnitude. When it comes to transposing analogue rights to AI-driven contexts, funding, expertise, data and hardware resources place companies in the driver's seat. They are in charge of translating antiquated, vague, user-oriented, analogue-era rights to a contemporary, technically specific

22 See https://perma.cc/G28B-VQAY.

23 In the US context, see *HUD*, *supra* note 2; the Biden administration has signalled its intent to review the rule, see The White House (2021).

24 For a critical view of trade secret privilege in criminal proceedings, see Wexler (2018).

25 With gratitude for an informal discussion in a seminar led by Jonathan Zittrain and Joichi Ito (2019).

26 See the section titled "Administrative Oversight" for the risk categories.

and corporate compliance-oriented context.[27] For government and administrative agencies deploying proprietary AI systems in forward-facing services, outsourcing also displaces public authority and expertise (Keats Citron 2008, 1296). In the context of online platforms moderating content, private adjudicative bodies such as the Facebook Oversight Board shape human rights norms faster than nation-states, treaty bodies and experts develop international human rights law (Douek 2021). Private companies can thus co-opt and bend the language of rights to serve their own interests. Similarly, companies at the helm of AI development can promote a "rights-lite" narrative that evacuates contestation while mainstreaming a thin conception of rights. Expecting companies to proactively adopt policies counter to their self-interest is at best idealistic; the remainder of this section offers pathways for policy makers to assert a more central role in the development of AI systems.

## Checks and Balances to Counter Industry Dominance

Policy makers can adapt the principles of constitutionalism to shape industry-driven AI development. At its core, constitutionalism distributes power to prevent the tyranny of one actor. The first component of constitutionalism consists of allocating power among separate branches of the government. Constitutionalism also sets up individual rights to further check against abuse by public actors.[28] These checks and balances encourage productive friction.[29] A common thread is introducing a plurality of interests to create a system of checks and balances.

Constitutionalism was originally deployed to check state power, but it can also guard against corporate hegemony. Digital constitutionalism can mitigate the prominence of industry in AI development, introducing a plurality of interests to influence the innovation agenda. This requires developing new governance structures for heavily decentralized contexts: "Traditional constitutionalism proceeds on the basis that the state is the most important source of power and, therefore, focuses on holding governments accountable. Digital constitutionalism must instead find a way to regulate power that is distributed among many actors within complex systems with many separate interacting components. For governments, this means radically rethinking how regulation can operate in a decentered government — where the state is not the only, or even the most powerful, actor seeking to regulate behavior" (Suzor 2019, 166). Policy makers should be intentional about distributing power among private actors to create an ecosystem conducive to various interests mutually checking one another. Co-regulation provides a concrete strategy to organize power among private actors to prevent abuse.

## Co-regulation

Co-regulation consists of governments and private actors sharing responsibility for drafting and enforcing standards (Hirsch 2011). It comes in many flavours. Policy makers can draw on industry guidelines to draft legislation, engage in negotiated rulemaking, or provide safe harbours to encourage industry codes of conduct. The following factors influence the optimal modality of co-regulation:

- whether the targeted industry consists of repeat players;
- the weight of reputation in a given sector;
- the internal organization of firms;
- an established compliance culture;
- the network between private actors; and
- the presence of sophisticated civil society players that could act as external checks (Kaminski 2019b, 1565–66).

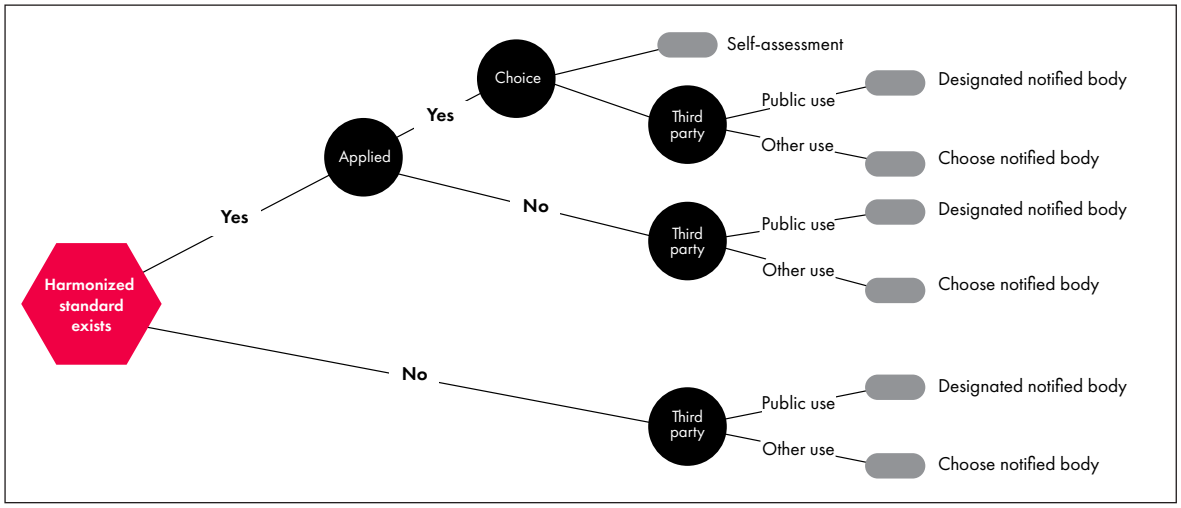
To note, some co-regulation modalities bleed into self-regulation. For example, regulators setting up rules and letting companies enforce them is arguably closer to self-regulation because enforcement is ultimately determinative.

27 See Lessig (2006, 166) on translating constitutional protections in new technological contexts; Waldman (2019, 627) on "shoehorning policy into codable algorithms"; Keats Citron (2008, 1268–71) on the mistakes, oversimplification and distortion of policy goals that result in tangible harms when programmers without deep policy expertise perform the front-line translation from policy to automated systems; Waldman (2021) on GDPR corporate compliance distorting policy goals from reducing consumer harm to handling corporate liability exposure.

28 See Santaniello et al. (2018, 324): "human rights are counter-institutions that embody the resistance of 'flesh-and-blood human beings against the structural violence of the matrix'" (citations and reference omitted).

29 In the vein of Montesquieu and Kant, James Madison (1788) suggests that "ambition must be made to counteract ambition." See also Weber (2019, 424).

## Figure 2: Conformity Assessment Procedure in the AI Act, Article 43




*Source:* Author.

## The EU AI Act

The European Union's AI Act prototype contains elements of co-regulation. For example, it requires producers of high-risk biometric identification systems to perform a pre-market conformity assessment.[30] The assessment essentially consists of verifying the producer's strategy for complying with the act. A quality management system ensures compliance with the act through safeguards such as quality control, data management, monitoring and incident reporting.[31] There are two pathways to assess conformity: internal control and third-party assessment[32] (see Figure 2 for decision flow).

The primary factor determining which pathway applies is whether the *Official Journal of the European Union* has published harmonized standards.[33] When such standards exist, the AI producer has the choice to apply them. If it elects to do so, article 43(1) lets it further choose between a streamlined self-assessment or a more comprehensive third-party assessment. Unless a public entity uses the biometric system in a specific area such as law enforcement or immigration, AI producers can resort to the third-party assessor of their choice, which the act calls a "notified body." If the producer declines to apply harmonized standards, or if such standards are unavailable, self-assessment is off the table and it must resort to a notified body for an external compliance assessment. Again, it can choose to do business with the notified body of its choice unless a public entity rolls out the system for specific purposes enumerated in the act.

A state-run accreditation system and investigative powers for the European Commission guarantee public control over the notified bodies.[34] Article 33 further sets out independence and impartiality requirements for notified bodies, not unlike the features of quasi-judicial bodies. Maintaining state-issued certification provides an additional check against notified bodies rubber-stamping biometric products, as they depend on certification to operate. The act therefore sets up a cottage industry of third parties assessing legislative compliance; notified bodies operate as for-profit administrative agencies that check biometric products' conformity with the substantive requirements of the act in articles 8–15.

The AI Act's conformity assessment procedure for biometric recognition products adapts the constitutional principle of checks and balances to structure power among private actors. The European Commission reinvents itself as the orchestrator of the overall power distribution between public and private actors. Oversight of the notified bodies and the option to endorse

30 *AI Act, supra* note 14, arts 19, 43(1), Annex 3.

31 Ibid., arts 17, 43.

32 Ibid., art 43.

33 Ibid., arts 40, 43. At the time of writing, it is unclear who will draft those standards. Should private entities draft them, the legislation will effectively incorporate privately drafted standards.

34 Ibid., arts 30, 33, 37.

harmonized standards maintain indirect legislative and judicial functions for public actors. AI producers perform the role of an executive branch, making ground-level compliance determinations and deciding on the pathway to ascertain said compliance. Notified bodies perform the quasi-judicial function of administrative agencies, reviewing the producer's determinations through the lens of their specialized, arm's-length expertise.

This arrangement embodies digital constitutionalism in the sense that the regulator structures relationships between private parties to foster checks and balances and lets them perform the last mile of compliance. Broad-stroke legislation sets high-level requirements and provides private actors with the discretion to tailor a pathway to conformity depending on their internal capacity and business model. Yet another way to describe this light-touch intervention is setting a default low-cost pathway through harmonized standards, but providing companies with the option to customize their compliance strategy through higher-cost third-party assessment. The flexibility pertaining to the use of standards, the type of assessment and the specific third party performing it illustrates the potential of co-regulation.

## Technical Standards

Another co-regulation approach consists in intervening on technical architectures through standards. Reached through compromise and common drafting, technical standards embody the consensus of experts about best practices. Regulators can hinge compliance with sector-specific laws on conformity with standards, effectively incorporating them by reference. Brazil, India, Singapore and the United Kingdom have already done so in other areas (Cihon 2019, 16, note 69). In the context of automated decision making and profiling, the advisory body to the GDPR suggests "obtaining contractual assurances for third-party algorithms that auditing and testing has been carried out and the algorithm is compliant with agreed standards,"[35] which could refer to technical standards (Kaminski 2019b, 1600, note 358).

Procurement rules for selling AI systems to public entities could require conformity with technical standards certified by third-party organizations, including the International Organization for Standardization,[36] the Institute for Electrical and Electronics Engineers,[37] and Canada's Chief Information Officer Strategy Council.[38] Policy makers can also participate directly in standards development. For example, they can assign personnel to observe or contribute to common drafting. As the UN Guiding Principles on Business and Human Rights reiterate, states have a duty to protect individuals in their territory against human rights abuses. This duty entails "appropriate steps to prevent, investigate, punish and redress private actors' abuse" (United Nations Human Rights Office of the High Commissioner 2011, 3, principle 1, commentary). Participation in standard making falls within the ambit of such preventive steps to ensure adequate protection against AI-driven harms, including the above-mentioned equality challenges.

Whether via the assessment procedure in the European Union's AI Act or via technical standards, co-regulation favours buy-in through an increased sense of ownership for private actors actively shaping the rules. Co-regulation can turn the adversarial dynamic between regulators and the regulated into a joint problem-solving enterprise (Hirsch 2011, 466–67). On the flip side, industry actors are unlikely to reveal detrimental information, and the opportunity for public participation is residual at best (ibid., 468).

Two additional downsides arise in the AI space. First, co-regulation favours incumbent companies with established government ties and hefty public policy resources. As a handful of dataopolies currently dominate the market (Analytics Insight 2021), this dynamic is bound to be particularly salient. Second, the asymmetry between public and private domain expertise remains stark. If public actors cannot partake in granular analysis of the topic at hand, nominal co-regulation can morph into actual self-regulation. This form of industry capture is especially problematic because collaborative drafting masks the respective input of industry and public actors. Compared to lobbying efforts subject to some transparency, coopting co-

35 EC, Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679,* 17/EN WP251rev.01 (2018) at 32, online: <https://ec.europa.eu/newsroom/article29/items/612053>; *GDPR, supra* note 10, art 22.

36 See www.iso.org/committee/6794475.html.

37 See https://sagroups.ieee.org/7003/.

38 See https://ciostrategycouncil.com/standards/101_2019/.

regulation in this fashion is even more opaque. It is harder for external stakeholders to challenge this subtle form of influence because the unverifiable assumption of meaningful public actor input lends a veneer of legitimacy to the process. Pushed to its extreme, this form of industry capture by a handful of powerful companies leads to "legal endogeneity" (Metcalf et al. 2021, 66)[39] whereby an institution constructs the meaning of law and applies it to itself. These concerns converge around the lack of effective checks on private power, the very issue co-regulation is supposed to solve. Absent careful consideration of these risks, co-governance efforts to alleviate industry power can actually accelerate it. Yet properly implemented, co-regulation can push back against the hegemony of industry actors across the AI space.

Given recent breakthroughs allowing machine learning to do more with less data, the well-trodden path of personal data protection becomes less effective to address AI-driven harms. As a result, the case for a systemic approach to AI governance is even more compelling. Chipping away at the governance question will not only foster a more predictable environment for business development but it will also open the door for longer-term speculative imaginaries. Stepping out of the industry-driven straitjacket, we can shift attention from mitigating harms to generating new possibilities for AI development.

## Conclusion

In light of their unique duties to the public good and capabilities to intervene systemically, policy makers need to assume a front-and-centre role for ensuring beneficial AI systems. This paper sought to offer them a fuller tool kit to bolster their role in AI governance. Opening the aperture allows us to recognize that both turning up the heat of intervention through traditional regulatory tools and pivoting toward new types of intervention are on the table. Each approach comes with benefits and downsides, such that effective AI governance should strategically combine them to maximize impact. It is not a zero-sum game where regulators pursue a single approach to the detriment of all others. Instead, they should deploy many instruments in their tool box, simultaneously coming at AI harms from different angles. Like the Swiss cheese model conveying the idea that we must layer different types of protection against the coronavirus disease 2019 pandemic, public AI governance demands a systemic approach conjugating many regulatory actions. Charting this course will require context-sensitive consideration of the interaction between the different strategies to prevent interference among them.

39 See also Kaminski (2019b, 1581) on private actors displacing courts in interpreting rights in algorithmic contexts.

# Works Cited


Altman, Micah, Alexandra Wood and Effy Vayena. 2018. "A Harm-Reduction Framework for Algorithmic Fairness." *IEEE Security & Privacy* 16 (3): 34–45. https://papers.ssrn.com/abstract=3225887.

Analytics Insight. 2021. "Top 5 Artificial Intelligence Companies with Market Share." Analytics Insight, March 27. www.analyticsinsight.net/top-5-artificial-intelligence-companies-with-market-share.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias." ProPublica, May 23. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila and Francisco Herrera. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115.

Black, Julia and Andrew D. Murray. 2019. "Regulating AI and Machine Learning: Setting the Regulatory Agenda." *European Journal of Law and Technology* 10 (3): 1–21.

Brown, Ian and Christopher T. Marsden. 2013. *Regulating Code: Good Governance and Better Regulation in the Information Age*. Cambridge, MA: The MIT Press.

Buolamwini, Joy and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.

Cave, James. 2016. "Is This A Muffin Or A Chihuahua?" HuffPost, March 10. www.huffpost.com/entry/muffin-or-chihuahua_n_56e05c3ee4b065e2e3d461f2.

Celis, L. Elisa, Vijay Keswani and Nisheeth K. Vishnoi. 2020. "Data preprocessing to mitigate bias: A maximum entropy based approach." *Proceedings of the 37th International Conference on Machine Learning* 119: 1349–59.

Cihon, Peter. 2019. *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Oxford, UK: University of Oxford.

Citron, Danielle Keats. 2008. "Technological Due Process." *Washington University Law Review* 85 (6): 1249–1313.

Cohen, Julie E. 2019. *Between Truth and Power: The Legal Constructions of Informational Capitalism*. New York, NY: Oxford University Press.

Corbett-Davies, Sam, Emma Pierson, Avi Feller and Sharad Goel. 2016. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, October 17. https://perma.cc/2WAM-7QPM.

Douek, Evelyn. 2020. "The Limits of International Law in Content Moderation." *UC Irvine Journal of International, Transnational, and Comparative Law* 6: 37–76.

Element AI and Nesta. 2019. "Data Trusts: A new tool for data governance." https://hello.elementai.com/rs/024-OAQ-547/images/Data_Trusts_EN_201914.pdf.

Ellul, Joshua, Stephen McCarthy, Trevor Sammut, Juanita Brockdorff, Matthew Scerri, Gordon J. Pace. 2021. "A Pragmatic Approach to Regulating Artificial Intelligence: A Technology Regulator's Perspective." arXiv:2105.06267: 1–23.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

European Commission Directorate-General for Justice and Consumers, Directorate D – Equality. 2015. *Reversing the burden of proof: Practical dilemmas at the European and national level*. Brussels, Belgium: European Commission. doi:10.2838/05358.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." *IEEE 5th International Conference on Data Science and Advanced Analytics:* 1–10.

Goodman, Bryce and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation.'" *AI Magazine* 38 (3): 50–57.

Green, Ben. 2018. "Putting the J(ustice) in FAT." Berkman Klein Center Collection, February 26. https://medium.com/berkman-klein-center/putting-the-j-ustice-in-fat-28da2b8eae6d.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi and Fosca Giannotti. 2019. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5): 1–42.

Haataja, Meeri and Joanna J. Bryson. 2021. "What costs should we expect from the EU's AI Act?" *SocArXiv*: 1–6. https://osf.io/preprints/socarxiv/8nzb4/.

Haeck, Pieter. 2021. "Ex-Google boss slams transparency rules in Europe's AI bill." Politico, May 31. www.politico.eu/article/ex-google-boss-eu-risks-setback-by-demanding-transparent-ai.

Hao, Karen. 2020. "A radical new technique lets AI learn with practically no data." *MIT Technology Review*, October 16. www.technologyreview.com/2020/10/16/1010566/ai-machine-learning-with-tiny-data.

Hirsch, Dennis D. 2011. "The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-Regulation?" *Seattle University Law Review* 34 (2): 439–65.

Hoefler, Torsten, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden and Alexandra Peste. 2021. "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks." arXiv:2102.00554.

Howard, Robert M. and Amy Steigerwalt. 2012. *Judging Law and Policy: Courts and Policymaking in the American Political System*. New York, NY: Routledge.

Joshi, Shalmali, Chirag Agarwal and Hima Lakkaraju. 2021. "Explainable ML in the Wild: When Not to Trust Your Explanations." Tutorial presented at the Conference on Fairness, Accountability, and Transparency, May 5. https://perma.cc/DXU7-3UQA.

Kaminski, Margot E. 2019a. "The Right to Explanation, Explained." *Berkeley Technology Law Journal* 34 (1): 189–218.

———. 2019b. "Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability." *Southern California Law Review* 92: 1529–1616.

Katyal, Sonia K. 2019. "Private Accountability in the Age of Artificial Intelligence." *UCLA Law Review* 66: 54–141.

Klein, Ezra. 2021. "Transcript: Ezra Klein Interviews Sam Altman." *The New York Times,* June 11. www.nytimes.com/2021/06/11/podcasts/transcript-ezra-klein-interviews-sam-altman.html.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Innovations in Theoretical Computer Science*: 1–23.

Koepke, John Logan and David G. Robinson. "Danger Ahead: Risk Assessment and the Future of Bail Reform." *Washington Law Review* 93: 1725–1807.

Kusisto, Laura and Mengqi Sun. 2021. "The Facebook Whistleblower, Frances Haugen: Does the Law Protect Her?" *The Wall Street Journal*, October 5. https://perma.cc/6L7T-VXQC.

Lessig, Lawrence. 2006. *Code: Version 2.0*. New York, NY: Basic Books.

MacCarthy, Mark. 2018. "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms." *Cumberland Law Review* 48 (67): 75–76.

Madison, James. 1788. "The Structure of the Government Must Furnish the Proper Checks and Balances Between the Different Departments from the New York Packet." *The Federalist Papers* No. 51, February.

Marchisio, Emiliano. 2021. "In support of 'no-fault' civil liability rules for artificial intelligence." *SN Social* Sciences 1 (54): 1–25.

Masnick, Mike. 2018. "Companies Respond To The GDPR By Blocking All EU Users." Techdirt, May 10. www.techdirt.com/articles/20180509/14021739811/companies-respond-to-gdpr-blocking-all-eu-users.shtml.

McStay, Andrew and Lachlan Urquhart. 2019. "'This time with feeling?' Assessing EU data governance implications of out of home appraisal based emotional AI." *First Monday* 24 (10): 1–16. https://firstmonday.org/ojs/index.php/fm/article/view/9457.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh and Madeleine Clare Elish. 2021. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts." In *ACM Conference on Fairness, Accountability, and Transparency*, March 3–10, virtual event, Canada. https://papers.ssrn.com/abstract=3736261.

Mueller, Benjamin. 2021. "How Much Will the Artificial Intelligence Act Cost Europe?" Center for Data Innovation, July 26. https://datainnovation.org/2021/07/how-much-will-the-artificial-intelligence-act-cost-europe/.

National Fair Housing Alliance. 2021. "Facebook Settlement." https://nationalfairhousing.org/facebook-settlement/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown.

Open Data Institute. 2019. *Data trusts: lessons from three pilots*. https://theodi.org/article/odi-data-trusts-report/.

Pasquale, Frank. 2016. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff and Sharad Goel. 2020. "A large-scale analysis of racial disparities in police stops across the United States." *Nature Human Behaviour* 4: 736–45.

Rudin, Cynthia. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1: 206–15.

Rudin, Cynthia, Caroline Wang and Beau Coker. 2019. "The age of secrecy and unfairness in recidivism prediction." arXiv:1811.00731: 1–46.

Santaniello, Mauro, Nicola Palladino, Maria Carmela Catone and Paolo Diana. 2018. "The language of digital constitutionalism and the role of national parliaments." *International Communication Gazette* 80 (4): 320–36.

Sartor, Giovanni and Francesca Lagioia. 2020. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. Brussels, Belgium: European Union. www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.

Scherer, Matthew U. 2016. "Regulating artificial intelligence systems: risks, challenges, competencies, and strategies." *Harvard Journal of Law & Technology* 29 (2): 353–93.

Shaer, Matthew. 2016. "The False Promise of DNA Testing." *The Atlantic*. June. https://perma.cc/AJH6-HK99.

Solow-Niederman, Alicia. 2019. "Administering Artificial Intelligence." *Southern California Law Review* 93: 633–96.

South, Jeff. 2018. "More than 1,000 U.S. news sites are still unavailable in Europe, two months after the GDPR took effect." NiemanLab, August 7. www.niemanlab.org/2018/08/more-than-1000-u-s-news-sites-are-still-unavailable-in-europe-two-months-after-gdpr-took-effect.

Sucholutsky, Ilia and Matthias Schonlau. 2020. "'Less Than One'-Shot Learning: Learning N Classes From M<N Samples." arXiv:2009.08449: 1–8.

Suzor, Nicolas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge, UK: Cambridge University Press.

*The Wall Street Journal*. 2021. "Facebook Whistleblower Frances Haugen Testifies." October 5. YouTube video, 1:10. www.youtube.com/watch?v=GoSPmqqKams.

The White House. 2021. "Memorandum on Redressing Our Nation's and the Federal Government's History of Discriminatory Housing Practices and Policies." Presidential actions, January 26. www.whitehouse.gov/briefing-room/presidential-actions/2021/01/26/memorandum-on-redressing-our-nations-and-the-federal-governments-history-of-discriminatory-housing-practices-and-policies/.

Tutt, Andrew. 2017. "An FDA for Algorithms." *Administrative Law Review* 69 (1): 88–123. www.jstor.org/stable/44648608.

United Nations Human Rights Office of the High Commissioner. 2011. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*. HR/PUB/11/04. https://perma.cc/5YF8-WWNW.

Wachter, Sandra, Brent Mittelstadt and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7 (2): 76–99.

Wachter, Sandra, Brent Mittelstadt and Chris Russell. 2021. "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI." *Computer Law & Security Review* 41: 105567.

Waldman, Ari Ezra. 2019. "Power, Process, and Automated Decision-Making." *Fordham Law Review* 88 (2): 613–32.

———. 2021. "Outsourcing Privacy." *Notre Dame Law Review* 96 (4): 194–210.

Weber, Max. 2019. *Economy and Society: A New Translation*. Cambridge, MA: Harvard University Press.

Wexler, Rebecca. 2018. "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System." *Stanford Law Review* 70 (5): 1343–429.

Zittrain, Jonathan and Joichi Ito. 2019. "The Ethics and Governance of Artificial Intelligence." https://opencasebook.org/casebooks/74426-the-ethics-and-governance-of-artificial-intelligence.

Centre for International
Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

@cigionline