Harvard Journal of Law & Technology
Volume 35, Number 1 Fall 2021

AN INSTITUTIONAL VIEW OF ALGORITHMIC IMPACT ASSESSMENTS

*Andrew D. Selbst**

ABSTRACT

Scholars and advocates have proposed algorithmic impact assessments ("AIAs") as a regulatory strategy for addressing and correcting algorithmic harms. An AIA-based regulatory framework would require the creator of an algorithmic system to assess its potential socially harmful impacts before implementation and create documentation that can be used later for accountability and future policy development. In practice, an impact assessment framework relies on the expertise and information to which only the creators of the project have access. It is therefore inevitable that technology firms will have an amount of practical discretion in the assessment, and willing cooperation from firms is necessary to make the regulation work. But a regime that relies on good-faith partnership from the private sector also has strong potential to be undermined by the incentives and institutional logics of the private sector. This Article argues that for AIA regulation to be effective, it must anticipate the ways that such regulation will be filtered through the private sector institutional environment.

This Article combines insights from governance, organizational theory, and computer science to explore how future AIA regulations

---

* Assistant Professor, UCLA School of Law. Thanks to Denise Anthony, Jane Bambauer, Hannah Bloch-Wehba, William Boyd, Ann Carlson, Julie Cohen, Michele Gilman, Kiel Brennan-Marquez, Isabelle Geczy, Jonathan Glater, Nik Guggenberger, Dennis Hirsch, Margot Kaminski, Pauline Kim, Alicia Solow-Niederman, Rory Van Loo, Andrew Verstein, Eugene Volokh, Ari Waldman, Steven Winter, Noah Zatz, and the participants of the Pepperdine Caruso School of Law Faculty Workshop, Max Planck Institute for Research on Collective Goods Seminar, University of Arizona James E. Rogers College of Law Faculty Workshop, Gershenson Faculty Workshop at Wayne State University Law School, Junior Law & Tech Scholars Workshop, the 2020 Privacy Law Scholars' Conference, and the UCLA Summer Faculty Workshop for helpful comments and insights on earlier drafts. Thanks especially to my former Data & Society colleagues Madeleine Elish, Mark Latonero, Jake Metcalf, Manny Moss, and Elizabeth Watkins for many conversations thinking through these issues together before this project coalesced into a paper. Thanks as well to Izabella Higson and Cecilia Bobbitt for invaluable research assistance, and the editors of the *Harvard Journal of Law and Technology* for their outstanding and professional work in preparing this Article for publication.

may be implemented on the ground. An AIA regulation has two main goals: (1) to require firms to consider social impacts early and work to mitigate them before development, and (2) to create documentation of decisions and testing that can support future policy-learning. The Article argues that institutional logics, such as liability avoidance and the profit motive, will render the first goal difficult to fully achieve in the short term because the practical discretion that firms have allows them room to undermine the AIA requirements. But AIAs can still be beneficial because the second goal does not require full compliance to be successful. Over time, there is also reason to believe that AIAs can be part of a broader cultural shift toward accountability within the technical industry. This will lead to greater buy-in and less need for enforcement of documentation requirements.

Given the degree to which an AIA regulation will rely on good faith participation by regulated firms, AIAs must have synergy with how the field works rather than be in tension with it. For this reason, the Article argues that it is also crucial that regulators understand the technical industry itself, including the technology, the organizational culture, and emerging documentation standards. This Article demonstrates how emerging research within the field of algorithmic accountability can also inform the shape of AIA regulation. By looking at the different stages of development and so-called "pause points," regulators can know at which points firms can export information. Looking at AI ethics research can show what social impacts the field thinks are important and where it might miss issues that policymakers care about. Overall, understanding the industry can make the AIA documentation requirements themselves more legible to technology firms, easing the path for a future AIA mandate to be successful on the ground.

TABLE OF CONTENTS

I. INTRODUCTION

In broad strokes, the arguments about the perils and promise of artificial intelligence ("AI") are well-rehearsed. AI can crunch quantities of data that no human can. It promises to find patterns that humans would otherwise miss; to be ever-vigilant where humans have to divide their time; and to be precise, mechanistic, and efficient where humans are arbitrary, sloppy, and biased.[1] AI also brings risks of harmful outcomes due to replication of human bias or other programmed-in biases;[2] errors that result from AI's ignorance of social

1. *See generally, e.g.*, Curtis E.A. Karnow, *The Opinion of Machines*, *in* THE CAMBRIDGE HANDBOOK OF THE LAW OF ALGORITHMS 16 (Woodrow Barfield ed., 2021).

2. *See, e.g.*, SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION 24 (2018); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 875 (2017).

and cultural contexts;[3] displacement of labor[4] and reduction of the tax base;[5] and difficulties of oversight stemming from a lack of transparency,[6] predictability,[7] and explainability,[8] as well as the transfer of decisionmaking authority from the democratic process to programmers.[9]

Given the power and great potential for harm that AI presents, legal scholars, policymakers, and advocates are looking to possible regulatory responses, including pre-existing remedies in anti-discrimination law,[10] administrative law or due process,[11] and tort law.[12] Scholars in other fields have been working too: looking to build

---

3. *See, e.g.*, Alistair Barr, *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*, WALL ST. J. (July 1, 2015, 3:41 PM), https://www.wsj.com/articles/BL-DGB-42522 [https://perma.cc/PVN2-THNY]; Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 59, 59 (2019).

4. *See generally* MARK MURO, ROBERT MAXIM & JACOB WHITON, BROOKINGS, AUTOMATION AND ARTIFICIAL INTELLIGENCE: HOW MACHINES ARE AFFECTING PEOPLE AND PLACES (2019), https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/ [https://perma.cc/XF53-6RDZ].

5. Matt Simon, *Who Will Pay for the Future if Not the Robots?*, WIRED (May 30, 2017, 7:00 AM), https://www.wired.com/2017/05/will-pay-future-not-robots/ [https://perma.cc/UQN4-XKNN].

6. *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1373–76 (2018); Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1236 (2019).

7. *See, e.g.*, Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots*, *in* ROBOT LAW 102, 107 (Ryan Calo et al. eds., 2016); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 365 (2016); Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 542 (2015).

8. Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1109 (2018); Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 842 (2018); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1341–42 (2020).

9. Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. (forthcoming 2021); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254 (2008).

10. Barocas & Selbst, *supra* note 2, at 694; Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1726–27 (2020); *see also, e.g.*, Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick & Jintong Tang, *The Law and Policy of People Analytics*, 88 U. COLO. L. REV. 961, 1010 (2017); Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1056 (2017); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 570 (2018); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 171 (2017); Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. (forthcoming 2021).

11. Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017); Citron, *supra* note 9, at 1178–79; Calo & Citron, *supra* note 9, at 820.

12. *E.g.*, Selbst, *supra* note 8, at 1320–22; W. Nicholson Price II, *Medical Malpractice and Black-Box Medicine*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 295, 300–01 (I. Glenn Cohen et al. eds., 2018); Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles:*

fairer, more interpretable,[13] and reviewable[14] systems; arguing for a better understanding of how algorithmic systems are situated in social contexts;[15] and advocating for public participation in algorithmic governance.[16]

But while AI's problems are recognized generally, many of the specifics are still not understood. We are still not able to predict in detail whether a particular AI is likely to be more or less biased than humans; what makes it so; and how the answers may vary across different contexts, such as policing, employment, credit, or public benefits. The public does not have insight into how specific decisions that firms make when designing or implementing AI systems affect their downstream results, or how — or, indeed, if — firms are measuring or addressing those impacts. We do not know how policy goals are translated into algorithmic systems, or the political choices that the algorithmic systems actually represent.[17] Because almost all AI systems — even those used in the public sector[18] — are developed privately and secretly,[19] the public knows very little about them.[20]

State Tort Liability, Automobile Insurance, and Federal Safety Regulation, 105 CALIF. L. REV. 1611, 1619 (2017); Curtis E.A. Karnow, *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, *in* ROBOT LAW 51 (Ryan Calo et al. eds., 2016); Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127, 145 (2019); Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 MICH. ST. L. REV. 1, 51 (2017); F. Patrick Hubbard, *'Sophisticated Robots': Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803, 1854 (2014); Kyle Graham, *Of Frightened Horses and Autonomous Vehicles: Tort Law and its Assimilation of Innovations*, 52 SANTA CLARA L. REV. 1241, 1269 (2012); Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321, 1326–28 (2012).

13. Kacper Sokol & Peter Flach, *Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 56, 65 (2020).

14. Jennifer Cobbe, Michelle Seng Ah Lee & Jatinder Singh, *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY TRANSPARENCY 598 (2021).

15. *E.g.*, Chelsea Barabas, Colin Doyle, JB Rubinovitz & Karthik Dinakar, *Studying Up: Reorienting the Study of Algorithmic Fairness Around Issues of Power*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 167, 170 (2020); Michael Katell et al., *Toward Situated Interventions for Algorithmic Equity: Lessons from the Field*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 45, 45 (2020); Francois Roewer-Despres & Janelle Berscheid, *Continuous Subject-in-the-Loop Integration: Centering AI on Marginalized Communities*, NEURIPS RESISTANCE AI WORKSHOP 1, 2–3 (2020).

16. Ngozi Okidegbe, *The Democratizing Potential of Algorithms?*, 54 CONN. L. REV. (forthcoming 2021) (noting the growing consensus while criticizing approaches that are insufficiently attentive to power dynamics).

17. Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 119 (2018).

18. *See id.* at 152.

19. *Id.* (noting that of all their requests only Allegheny County was fully responsive). Allegheny County's approach to its child welfare system has become a famous and oft-studied example precisely because the level of transparency is so rare. *See, e.g.*, Alexandra Chouldechova, Emily Putnam-Hornstein, Diana Benavides-Prado, Oleksandr Fialko &

For this reason, one regulatory approach that has gained favor in recent years is regulation requiring Algorithmic Impact Assessments ("AIAs").[21] The impact assessment approach has two principal goals. The first is to get the people who build systems to think methodically about the details and potential impacts of a complex project before its implementation, thereby heading off risks before they become too costly to correct.[22] As proponents of values-in-design have argued for decades, the earlier in project development that social values are considered, the more likely that the end result will reflect those social values.[23] The second goal is to create and provide documentation of the decisions made during development and their rationales, which in turn can lead to better accountability for those decisions and useful information for future policy interventions.

Since the passage of the National Environmental Policy Act ("NEPA") in 1969,[24] impact assessments have been a commonly replicated tool, used in a wide variety of contexts: environmental,[25] sentencing,[26] privacy,[27] human rights,[28] data protection,[29] police

---

Rhema Vaithianathan, *A Case Study of Algorithm-assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 134, 134–35 (2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 5 (2018); Dan Hurley, *Can an Algorithm Tell When Kids Are in Danger?*, N.Y. TIMES MAG. (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html [https://perma.cc/S49Y-M4HQ].

20. Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773, 778 (2019) ("[G]overnment agencies purchasing and using these [machine learning] systems most often have no input into — or even knowledge about — their design or how well that design aligns with public goals and values."). We don't even really know what the definition of AI is! *See generally* Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORNELL L. REV. 287, 293–94, 357 (arguing that we lack definitions of robots and AI, that such definitions miss the point, and that we should regulate based on behavior).

21. *See infra* Part II.

22. In this Article, I discuss AIAs as private sector regulation. Impact assessments may seem a more natural fit in the public sector, as that is where they originated, but the analysis of public sector AIAs necessitates a different discussion entirely.

23. *See, e.g.*, Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANSACTIONS ON INFO. SYS. (TOIS) 330, 343–44 (1996); Ann Cavoukian, Scott Taylor & Martin E. Abrams, *Privacy by Design: Essential for Organizational Accountability and Strong Business Practices*, 3 IDENTITY & INFO. SOC'Y 405, 406 (2010).

24. National Environmental Policy Act of 1969, 42 U.S.C. §§ 4331–47.

25. *See, e.g.*, *id.*

26. *See* Jessica Erickson, Comment, *Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System*, 89 WASH. L. REV. 1425, 1444–45 (2014).

27. *See, e.g.*, E-Government Act of 2002, Pub. L. No. 107–347, 116 Stat. 2899 (2002).

28. *See generally* RORY MUNGOVEN, WALKING THE TALK EXPLORING METHODOLOGIES AND APPLICATIONS FOR HUMAN RIGHTS IMPACT ASSESSMENT BY THE UNITED NATIONS (2016).

29. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, 2016 O.J. (L 119) 1.

technology,[30] surveillance,[31] and — in Canada, where the AIA is already a reality — algorithmic decisionmaking.[32] They are used extensively at all levels of government.[33] And although NEPA originally intended impact assessments for the public sector, because the law was held to apply to any project that requires federal funding or permitting, the private sector has been conducting them for just as long as governments.[34] In the decades since NEPA's enactment, a field of "social impact assessment" has arisen with the aim of developing impact assessment methods and methodologies within the private sector.[35]

Impact assessments are most useful when projects have unknown and hard-to-measure impacts on society, when the people creating the project are the ones with the knowledge and expertise to estimate its impacts but have inadequate incentives to generate the needed information, and when the public has no other means to discern that information.[36] The AIA is attractive because we are now in exactly such a

---

30. *See Community Control Over Police Surveillance*, ACLU, https://www.aclu.org/issues/privacy-technology/surveillance-technologies/community-control-over-police-surveillance [https://perma.cc/MS8V-DQSY] (describing the success of the ACLU's model impact assessment law for police technologies); *see also* Laura M. Moy, *A Taxonomy of Police Technology's Racial Inequity Problems*, 2021 U. ILL. L. REV. 139, 176–81 (proposing "police technology equity impact assessments").

31. David Wright & Charles D. Raab, *Constructing a Surveillance Impact Assessment*, 28 COMPUT. L. & SEC. REV. 613, 616 (2012).

32. GOVERNMENT OF CANADA, ALGORITHMIC IMPACT ASSESSMENT (AIA), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html [https://perma.cc/HKL4-RWUN] [hereinafter CANADIAN AIA]. Though implemented as a requirement, there appears to be some question as to how frequently the AIA is completed when called for. *See* Tom Cardoso & Bill Curry, *National Defence Skirted Federal Rules in Using Artificial Intelligence, Privacy Commissioner Says*, GLOBE & MAIL (Feb. 8, 2021), https://www.theglobeandmail.com/canada/article-national-defence-skirted-federal-rules-in-using-artificial/ [https://perma.cc/L2CA-U3FU].

33. Bradley C. Karkkainen, *Toward A Smarter NEPA: Monitoring and Managing Government's Environmental Performance*, 102 COLUM. L. REV. 903, 905–06 (2002) (footnotes omitted) ("NEPA is without question the most widely emulated of the major U.S. environmental laws. It has inspired dozens of 'little NEPAs' at the state and local levels, numerous progeny around the globe, and countless imitators in other fields.").

34. *See* David J. Hayes & James A. Hourihan, *NEPA Requirements for Private Projects*, 13 B.C. ENVTL. AFFS. L. REV. 61, 61–62 (1985).

35. *See, e.g.*, William R. Freudenburg, *Social Impact Assessment*, 12 ANN. REV. SOCIO. 451, 451 (1986); Nicholas Diakopoulos et al., *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*, FAIRNESS, ACCOUNTABILITY & TRANSPARENCY IN MACH. LEARNING, https://www.fatml.org/resources/principles-for-accountable-algorithms [https://perma.cc/3MPV-23TA]; Ana Maria Esteves, Daniel Franks & Frank Vanclay, *Social Impact Assessment: The State of the Art*, 30 IMPACT ASSESSMENT AND PROJECT APPRAISAL 34, 34 (2012).

36. *See* Winter v. Nat. Res. Def. Council, Inc., 555 U.S. 7, 23 (2008) ("Part of the harm NEPA attempts to prevent in requiring an [environmental impact statement] is that, without one, there may be little if any information about prospective environmental harms and potential mitigating measures.").

situation with respect to algorithmic harms.[37] The public knows that there are potential harms associated with algorithmic systems but has neither the information nor the expertise to get into the weeds and discover what types of decisions in system design lead to particular types of problems. It will be difficult to address algorithmic harms more concretely or thoroughly without such information.

While AIAs may be a sound regulatory strategy in principle, a practical challenge arises when we consider that they will necessarily be implemented by the very firms building algorithmic technology.[38] The expertise and information contained within the industry itself is necessary for successful assessment of harms, and therefore the industry will have a hand in its own governance. This fact has certain consequences for the efficacy of the regulation in practice. Those consequences, and how to mitigate or address them, are the subject of this Article. It is necessary to understand the institutional forces at play in the organizations where systems will be built and impacts will be assessed. Only by understanding how the law is likely to be shaped and understood on the ground can we hope to use it to its fullest effect. This Article will argue in part that, once filtered through the institutional logics of the private sector, the AIA's first goal — to improve systems through better design — will only be effective in those organizations motivated by social obligation rather than mere compliance. However, the second goal — to produce the information needed for better policy and public understanding — is what really can make an AIA regime worthwhile, regardless of organizations' motivations.

That the current environment lends itself to an AIA approach does not mean that in a vacuum AIAs would be the most effective regulation of algorithmic systems possible. Quite the contrary. As this

---

37. *See* Michael Guihot, Anne F. Matthew & Nicolas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 456 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 169 (2017); Mulligan & Bamberger, *supra* note 20, at 778.

38. Another possibility I do not expressly consider here is that an ecosystem of private third-party independent assessors will arise. This may change the incentives but is far from a panacea. These professional assessors are likely to be financially beholden to the industry actors that they assess and will rely on cultivating a reputation of being friendly to the private sector actors they oversee, leading to a possible merging of incentives. There is some evidence that this is starting to happen already in the unregulated "algorithm auditing" space, where some firms are proposing "collaborative audits," and companies are co-opting the audit process for their own public relations purposes. Mona Sloane, *The Algorithmic Auditing Trap*, MEDIUM (Mar. 17, 2021), https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d [https://perma.cc/R9AD-KC9J] (quoting Wilson et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1, 10 (2021)); *see also* Alfred Ng, *Can Auditing Eliminate Bias from Algorithms?*, THE MARKUP (Feb. 23, 2021, 8:00 AM), https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms [https://perma.cc/6JJT-QMH3].

Article will detail, AIA regimes will likely not be effective enough to be the final word on policy. But, given the information disparities between developers on the one hand, and policymakers and the public on the other, regulation that can slow down the development process, create pathways for public input, and push information out to the public can be an important step toward both mitigating current harms and developing better, more concrete regulation in the future. There are certainly reasons to think we should skip this step entirely and immediately move toward more aggressive regulation. Such an approach may be called for in certain contexts, such as facial recognition, in which algorithmic systems pose unique dangers.[39] Additionally, as a matter of politics, reformers may get only one bite at the apple, which suggests that AIAs or any other stopgap regulation would in fact be a mistake.[40] These are important points, but it is also true that any regulation enacted without the information that an AIA regime would produce would be operating partly in the dark and therefore would result in certain unintended consequences and likely greater resistance from industry. Such a move might be the right one in the end, either because of the politics or for reasons that this Article discusses, including that the private sector can seriously undermine regimes of collaborative governance.[41] The aim of this Article, however, is to take seriously the practical reality that the private sector will be involved in any AIA regulation. If we are to decide whether AIAs are a good idea at all, or in case legislators move forward with the AIA idea as an achievable second-best approach, it will be important to understand what the best version of an AIA looks like.

The Article proceeds in four Parts. Part II introduces the AIA and explains why it is likely a useful approach. It offers three representative examples of algorithmic harm that have surfaced in scholarly literature and popular discourse: the biased hiring algorithm, the unexplained credit denial, and the unsafe medical AI. Each of these are real cases that implicate recognized algorithmic harms: discrimination, arbitrary decisionmaking, and physical injury. Part II demonstrates how current mechanisms of accountability for the relevant harm are difficult to apply, specifically due to the lack of knowledge the public has about the development processes. This is why it is nec-

---

39. Woodrow Hartzog & Evan Selinger, *Facial Recognition is the Perfect Tool for Oppression*, MEDIUM (Aug. 2, 2018), https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66 [https://perma.cc/XZY8-FQZD].

40. This is a serious concern, and I do not intend to gloss over it. The aim of this Article is to consider the idea of AIA regulation and how it might or might not be made effective when we consider private sector implementation. Even in the best case, though, I think it is likely a second-best approach in the long term. For someone convinced we really do only get one shot, it might not be a good approach at all.

41. *See* Ari Ezra Waldman, *Privacy, Practice, and Performance*, 110 CALIF. L. REV. (forthcoming 2021).

essary for regulation to focus on knowledge development before more substantive regulation can issue at a later time.

Part III briefly surveys different models of AIAs that have been proposed, as well as two alternatives: self-regulation and audits. These oversight mechanisms share many aspects but differ in important ways. Attending to these differences in light of the AIA's two regulatory goals, the Part discusses three factors that make the AIA a distinctive proposal from other types of constraint: timing, open-ended questions, and the need to access the resulting documentation.

Part IV examines how institutional forces shape regulation and compliance, seeking to apply those lessons to the case of AIAs. It draws on three theoretical frameworks to illustrate how an AIA regulation will be shaped in practice by institutional forces. First, because good faith cooperation by the private sector is required, this approach to AIAs fundamentally sits in the category of "collaborative governance" approaches to regulation. There is a vast legal literature on the benefits and drawbacks of such approaches, all of which will apply here. The second important framework is "legal managerialism," which observes that entrusting firms with policy implementation causes policy to be corrupted by the market logics that dominate the firms, and that having regulator and regulated work in cooperation may undermine regulatory goals. Thus, any collaborative regime must have safeguards in place and perhaps limits on what it can expect to accomplish on an individual firm level. Third, research has demonstrated that for various reasons some firms go beyond strict compliance. Combined with the neoinstitutional school of sociology's concept of "institutional isomorphism," which holds that firms in an industry tend to follow each other's practices, this suggests that an industry's norms can potentially be reshaped by just a few industry leaders. Part IV concludes that AIAs may not be fully successful in their first goal of getting individual firms to consider social problems early on, but that the second goal of policy-learning may be more successful because it does not require full substantive compliance.

Finally, Part V looks at what we can learn from the technical community. Once we abandon the idea of full compliance in a top-down regulatory regime and — for better or worse — embrace private industry as a partner, the compliance question shifts from enforcement to encouragement. This suggests that making compliance easier is of paramount importance, and that regulation can and should be designed with production processes and cycles in mind. This Part discusses many relevant developments within technology industry and scholarship: empirical research into how firms understand AI fairness and ethics; proposals for documentation standards coming from academic and industrial labs, trade groups, and standards organizations; and various self-regulatory framework proposals. To be successful,

AIA legislation should draw heavily on the emerging accountability frameworks coming from within computer science and the technology industry, and this Part offers suggestions for how to incorporate those frameworks in a regime of regulatory oversight.

II. ALGORITHMIC HARMS AND LIABILITY REGIMES

An Algorithmic Impact Assessment is a process in which the developer of an algorithmic system aims to anticipate, test, and investigate potential harms of the system before implementation; document those findings; and then either publicize them or report them to a regulator.[42] This technique has its roots in environmental law, in which the National Environmental Policy Act ("NEPA") imposed the requirement to document the choices made in project development, and the rationales for them, in an environmental impact statement ("EIS").[43] Since the passage of NEPA, the impact assessment has grown in prominence and is now used at every level of government[44] and in many different contexts, such as sentencing, privacy, human rights, data protection, police technology, and surveillance.[45]

As a regulatory approach, requiring firms to conduct impact assessments seeks both to change design processes to prevent or mitigate harm and to produce knowledge. The algorithmic accountability literature is replete with well-known examples of algorithmic harm to equality, dignity, autonomy, and safety.[46] The understanding that algorithmic decisionmaking can be harmful in all these ways has led to calls for accountability in general,[47] and the conversation has progressed to debating what type of regulation to implement and how.

So why impact assessments? Why discuss a regime focused on reforming design processes and producing knowledge rather than addressing the harms directly? Part of the answer is that right now there is just a lot we don't know. Due partly to the reflexive opacity of

---

42. *See* Selbst, *supra* note 37, at 169–82; Margot E. Kaminski & Andrew D. Selbst, *The Legislation That Targets the Racist Impacts of Tech*, N.Y. TIMES (May 7, 2019), https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html [https://perma.cc/57YU-5DRM] (emphasizing the need to report findings to the public or regulators).

43. 42 U.S.C. § 4332.

44. Bradley C. Karkkainen, *supra* note 33, at 905.

45. *See supra* notes 25–32.

46. *See, e.g.*, Ben Green & Salomé Viljoen, *Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 19, 19 (2020) (collecting sources).

47. *See e.g.*, ADA LOVELACE INST., AI NOW INST. & OPEN GOV'T P'SHIP, ALGORITHMIC ACCOUNTABILITY FOR THE PUBLIC SECTOR: LEARNING FROM THE FIRST WAVE OF POLICY IMPLEMENTATION 3–4 (2021), https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf [https://perma.cc/M82P-WP8M] (analyzing the "first wave" of algorithmic accountability policy interventions); *see also id.* at 57 (listing the interventions in the study).

technology companies and widespread claims of trade secrecy,[48] researchers are left trying to reconstruct what companies are doing from vague public statements.[49] But even aside from intentional secrecy, algorithmic systems are complex and have impacts that are difficult to understand and anticipate. As in the environmental context, in which impact assessments were pioneered, some of the harms are knowable but take resources and expertise to discover. The precise mechanisms for the known algorithmic harms are not yet understood and the ways to prevent or mitigate the harms are not obvious. Perhaps one day more straightforward regulation of specific harmful mechanisms will be appropriate, but a regime of documentation and knowledge production is necessary before we can get to that point, and some forced introspection and disclosure on the part of the producers of the harms can help in the meantime.

To demonstrate this, before getting into the substance of the AIA requirement, this Part reviews three examples of algorithmic harm where existing liability regimes fail to hold the creators of the harm to account, specifically because of a lack of knowledge about the development process. The examples below include a discriminatory hiring algorithm, an unexplained denial of a loan, and an unsafe medical diagnostic device. They are representative of three commonly discussed harms — discrimination, procedural injustice, and physical injury.

A. The Discriminatory Hiring Algorithm

Discrimination is easily the most recognized and discussed harm of algorithmic decisionmaking. Biased results that harm members of protected classes are a concern wherever algorithms are used to allocate opportunities. This includes employment,[50] credit,[51] housing,[52]

48. Wexler, *supra* note 6, at 1355; Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 99 (2019).

49. *See, e.g.*, Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1, 5 (2020) (analyzing industry practices for mitigating algorithmic bias based on public statements of companies).

50. Barocas & Selbst, *supra* note 2, at 674; Kim, *supra* note 2, at 861; Ajunwa, *supra* note 10, at 1692.

51. Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 148 (2016); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 1 (2014); Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375, 1379 (2014).

52. Virginia Foggo & John Villasenor, *Algorithms, Housing Discrimination, and the New Disparate Impact Rule*, 22 COL. SCI. TECH. L. REV. 1, 1 (2020); Shivangi Bhatia, *To 'Otherwise Make Unavailable': Tenant Screening Companies' Liability Under the Fair Housing Act's Disparate Impact Theory*, 88 FORDHAM L. REV. 2551, 2551 (2020); Valerie Schneider, *Locked Out by Big Data: How Big Data, Algorithms, and Machine Learning May Undermine Housing Justice*, 52 COLUM. HUM. RTS. L. REV. 251, 251 (2020).

policing,[53] pre-trial detention,[54] and sentencing,[55] among other areas. The concern for harm resulting from bias is the same concern that is examined as "fairness" within the computer science wing of algorithmic accountability discourse,[56] organized primarily around the annual "Fairness, Accountability, and Transparency" conference.[57] The example of a discriminatory hiring algorithm is, in turn, the most frequently discussed example of a discriminatory harm.

Bias in a hiring model can come from a number of different factors. Because a computer cannot know what makes a "good employee," the problem has to be reframed as an optimization problem that a computer can solve and render into a prediction.[58] Examples include predicting sales figures, tenure at a company, or scores on performance reviews. The choice of this optimization criterion is inherently subjective, and many versions are likely defensible, but they will lead to disparate results along demographic lines.[59] There is also subjectivity in how training data is collected and labeled. Which datasets are acquired or used will absolutely change the result, and any mismatch between the demographics of the training data and the population it applies to will lead to biased outcomes, as will any biases in the labeling of the training data, such as performance review scores — which are known to be biased.[60]

Despite any disproportionate effects on protected classes, it may be difficult for a plaintiff to win a Title VII suit against an employer who decides to use a competently designed machine learning model to predict their best candidate.[61] This is true for a few reasons, but the most important one rests on a single, high-level point. Anti-discrimination law is, as a practical matter, fault-based.[62] It is not

---

53. Selbst, *supra* note 37, at 121–23; ANDREW GUTHRIE FERGUSON, THE RISE OF BIG DATA POLICING: SURVEILLANCE, RACE, AND THE FUTURE OF LAW ENFORCEMENT 3 (2017).

54. Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2218 (2019); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1072 (2019).

55. Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 803 (2014); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 59 (2017); Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/94QT-DSEV].

56. *See, e.g.*, Xiang, *supra* note 10, at 6.

57. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY, https://facctconference.org [https://perma.cc/H595-7XXC].

58. Barocas & Selbst, *supra* note 2, at 679.

59. *Id.* at 680.

60. *Id.* at 680–88. There are still other ways a model can be biased, but these will suffice for now.

61. *See generally* Barocas & Selbst, *supra* note 2.

62. *See, e.g.*, Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 42–45 (2006); Tristin K. Green, *A Structural Approach as Antidiscrimination Mandate: Locating Employer Wrong*, 60 VAND. L. REV. 849, 871 (2007); George Rutherglen, *Ricci v Destefano: Affirmative Action and the Lessons of Adver-*

enough to say that an employee faced an adverse result because of their protected class; such result must have been attributable to the employer's actions.[63] This is obviously true of disparate treatment, with its intent standard, but it is equally true of disparate impact. While disparate impact doctrine is purportedly effects-based, the effects test is only the first step of a disparate impact analysis. The next steps — the business necessity and least discriminatory means tests — amount to other ways of asking whether the decisionmaker was at fault for the disproportionate impact.[64] Thus, if the relevant comparison is between an employer predicting future outcomes with a working algorithmic model and any other model that is inferior, such as traditional subjective interviewing, then the defendant is likely to win despite any discriminatory outcomes from the algorithmic model.

This initial conclusion assumes that the model is less discriminatory than non-algorithmic means of hiring, and that the choice is between this model and no model. Let us now tweak the hypothetical by assuming still that the model is less discriminatory than not using any model, but that there exists a hypothetical algorithmic model that is less discriminatory.[65] The business necessity and least discriminatory means tests would then come down in principle to whether the failure to build, use, or acquire a better version of the model renders the employer at fault.[66]

The need for AIAs becomes clear when we consider how to address this question. As with all fault-based questions, this inquiry is about whether the justifications for various decisions are broadly ac-

---

sity, 2009 SUP. CT. REV. 83, 98–99 (2009) ("An employer who has engaged in prohibited discrimination has done something wrong. Why else should liability be imposed upon that employer rather than someone else? The extended sense of 'discrimination' in the law dispenses with the need to prove intent, but it does not dispense with the need to prove fault."); Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 773–74 (2006) ("Without an element of blameworthiness, there is no basis on which to require remedial action . . . . [O]nce the Supreme Court moved away from an immediate locus of blame, it had an increasingly difficult time assigning liability or requiring remedial action.").

63. *See* Noah D. Zatz, *Disparate Impact and the Unity of Equality Law*, 97 B.U. L. REV. 1357, 1397 (2017) (distinguishing the discrimination injury from employer responsibility for it).

64. *See* Alan David Freeman, *Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine*, 62 MINN. L. REV. 1049, 1053–57 (1978) (discussing the "perpetrator perspective").

65. *See* Andrew D. Selbst, Suresh Venkatasubramanian & I. Elizabeth Kumar, *The Legal Construction of Black Boxes* (unpublished manuscript) (on file with author) (discussing courts' need to consider counterfactual versions of algorithmic systems).

66. The language of Title VII asks whether an employer "refuses" to use a less discriminatory model, seemingly putting a thumb on the scale for the defendant, but the word refuses has never been clarified. Following David Oppenhemier's observations that discrimination law has hidden aspects of negligence, *see generally* David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 899 (1993), perhaps what is meant is that the employer *unreasonably* fails to use a better version of the model. *See* Barocas & Selbst, *supra* note 2, at 710–11.

ceptable. Why was the specific optimization criterion chosen? Were others tested? Why was a certain training set chosen? Were others available? Did they cost too much? If an algorithmic system was purchased off the shelf, were alternatives sought and tested? Why were they rejected? These are the types of questions that would enable a jury or judge to decide whether the employer was at fault for a failure to have a less discriminatory alternative.[67] Right now, there is no regulatory incentive to do the testing or write the answers to those questions down, and these are exactly the sort of design questions that would be addressed in an AIA. If we want Title VII to be effective, we need documentation to exist — the same documentation that would constitute the AIA. The ex ante documentation requirement enables the possibility of an ex post remedy.

Thus, there is a procedural reason to have AIAs: to provide the information necessary for Title VII and other existing discrimination law to work better. But there is a substantive reason as well. A judgment of liability is a binary question. A person is either at fault or not; a plaintiff wins or a defendant does. In such a high-stakes environment, plaintiffs often lose, algorithm or no algorithm. There is no general agreement about when certain subtle choices that lead to disproportionate results might constitute wrongful discrimination.[68] Where there is uncertainty about the moral or legal valence of small actions with big, but perhaps unpredictable effects, and where an employer has adopted a model that improves over the status quo, courts may be hesitant to label "good" people as discriminators.[69]

Once discrimination law is set up to be defendant-friendly, there is little incentive for mitigation of disparate impacts. But there is good reason to believe that mitigation is a more worthwhile goal. If the higher-order goal of anti-discrimination is to create a more equal society, then judging individual employer fault is a rather high-stakes and indirect way of getting there. Because a lot of what algorithmic models do is reflect any number of structural and institutional factors that have already led to fewer opportunities for oppressed classes of people, "society" is likely to be seen as more to blame than individual employers.

This point is arguably even more fundamental to the conflict between machine learning systems and discrimination. As Issa Kohler-

---

67. Of course, how we determine whether a given choice that leads to a disproportionate outcome is in fact wrongful is a whole separate challenge, but in the long-term, perhaps one that can likely be sorted out on a case-by-case basis. *See* Bagenstos, *supra* note 62, at 34–36; Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 460 (2001).

68. *See* Bagenstos, *supra* note 62, at 34–36; Sturm, *supra* note 67, at 460. *See generally* George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313 (2006).

69. *See, e.g.*, Selmi, *supra* note 62, at 773–74.

Haussman has argued, if you accept that protected class status — especially race — is socially constructed rather than biologically derived, then counterfactual reasoning about it does not work.[70] If what it means to be Black in the United States is not based on skin color, but is instead a condition of Blackness that comprises all the historical disadvantages that that entails — fewer opportunities at every stage of life, trauma built up over time that may impede the ability to take advantage of opportunity that does present itself, lower average generational wealth and life expectancy, worse environmental conditions — then simply changing the race variable in a machine learning model to encode a different skin color does not accurately account for the counterfactual. A person coded white in the model also has that whiteness reflected in the compounded advantages that show in the very data we use to train the model: better educational outcomes, more opportunities for experience, and so on. While this point is similar mechanically to the observation that due to historical discrimination, a model may in some sense be both "accurate" and "discriminatory" because it is picking up on patterns of structural discrimination latent in society, Kohler-Hausmann would argue that the model is just describing race itself, or at least is partially doing so, and the description of "accurate-but-with-disparate-impact" is incoherent.[71] The only way to properly generate a counterfactual that could be used to measure the specific impact of race would be to remove the racial component from all variables that have it. But if we could do that, we would not need to build machine learning systems to predict outcomes, because we would already know the degree to which each variable contributes to the answer, assuming there are any objective answers to be found in the first place. Attempts to counteract the effects of race in a model — in effect aiming to predict race — will therefore not be fruitful, and it is more worthwhile instead to focus on decreasing the substantive impact of race on decisions proactively. Thus, by connecting upstream decisions to downstream impacts, AIAs can lead developers to focus on mitigation, which may be more beneficial than attempting to regulate discrimination through liability rules.

B. The Unexplained Loan Denial

A second common example of algorithmic harm is the unexplained adverse result, the fate sealed by an inscrutable black box.[72]

---

70. Issa Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination*, 113 NW. U. L. REV. 1163, 1163 (2019).

71. *Id.* at 1172.

72. *E.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 1–18 (2016).

The most frequent example is a denial of a loan,[73] but other benefits, such as unemployment insurance[74] or Medicaid disbursements,[75] also fall into this category. In the European Union's General Data Protection Regulation ("GDPR"), the provisions recognizing this harm are quite broad indeed, requiring "meaningful information"[76] about the logic of processing for any "legal" or "similarly significant" effects.[77]

The need for explanations of algorithmic systems is a consequence of the secrecy, opacity, complexity, inscrutability, and non-intuitiveness of algorithmic decisionmaking.[78] Algorithms crunch numbers to find patterns, which then become decision rules written into code. These decision rules are so complex that even the most transparent system would still run into the problem that a person who read through it and traced every decision could still not understand it as a whole.[79]

There are several reasons that a failure to have an explanation for a denial can be an algorithmic harm. An adverse credit determination is a good example from the private sector, but the harm from a failure to explain is much more general.[80] Explanations are necessary to respect a person's dignity and autonomy. Daniel Solove has argued that the best literary analogy for our information era is Franz Kafka's *The*

---

73. *E.g.*, Citron & Pasquale, *supra* note 51, at 1.

74. *See* JULIA SIMON-MISHEL, MAURICE EMSELLEM, MICHELE EVERMORE, ELLEN LECLERE, ANDREW STETTNER & MARTHA COVEN, THE CENTURY FOUNDATION, CENTERING WORKERS—HOW TO MODERNIZE UNEMPLOYMENT INSURANCE TECHNOLOGY 2 (2020), https://tcf.org/content/report/centering-workers-how-to-modernize-unemployment-insurance-technology [https://perma.cc/CP6S-UFWM].

75. *See* LYDIA X.Z. BROWN, MICHELLE RICHARDSON, RIDHI SHETTY & ANDREW CRAWFORD, CENTER FOR DEMOCRACY & TECHNOLOGY, CHALLENGING THE USE OF ALGORITHM-DRIVEN DECISION-MAKING IN BENEFITS DETERMINATIONS AFFECTING PEOPLE WITH DISABILITIES 6 (2020), https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/ [https://perma.cc/DPH9-882Q].

76. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, art. 15(1), 2016 O.J. (L 119) 43; *see also* Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 233 (2017).

77. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, art. 22(1), 2016 O.J. (L 119) 46.

78. Selbst & Barocas, *supra* note 8, at 1089–99.

79. *See id.* at 1094–96.

80. *See generally* Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851 (2019). Public benefits denials implicate similar failure-to-explain harms as loan denials, but they are treated quite differently in the U.S. because the public sector is subject to constraints arising from due process and administrative law. Even given the solicitude that administrative law pays to explanatory needs, the legal status of inscrutable algorithmic decisionmaking is not clear in the public sector. While due process requires *some* degree of transparency, *see* Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 41 (2019), automated decisionmaking inherently blurs the line between adjudication and rule, rendering a challenge of a specific denial likely impotent. *See* Citron, *supra* note 9, at 1249 ("[A]utomated decision making systems combine individual adjudications with rulemaking while adhering to the procedural safeguards of neither.").

Trial, because it best captures the dehumanization of unexplained decisions based on data.[81] Legally, lack of explanation could be considered a denial of procedural justice, the notion that a system of law should intrinsically respect people's dignity and autonomy.[82] Other reasons for explanation are instrumental rather than intrinsic.[83] One is that explanation enables people to regulate their behavior.[84] There are many websites dedicated to teaching people how to raise their credit scores because people do not understand them but want to behave in ways that will enable them to receive a loan in the future. Explanations allow that kind of corrective action. Finally, explanations offer a basis for evaluation and recourse.[85] A person cannot know whether the algorithmic model has violated some right of hers, and subsequently contest it, unless she gets an explanation about what the denial was based on. The need to contest a decision is accordingly one of the animating purposes for the hotly debated "right to explanation" in the GDPR.[86]

There are many approaches to explanation, and no single approach can achieve every goal. An engineer troubleshooting a model wants to know about general trends and if the model fails on certain subsets of cases with regularity; the engineer is not at all interested in individual outcomes. The consumer who is denied credit, however, wants to know why she, specifically, was denied credit, and the regu-

---

81. Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1419–30 (2001).

82. *See* Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181, 262–63 (2004); *id.* at 183 ("[P]rocedural justice is deeply entwined with the old and powerful idea that a process that guarantees rights of meaningful participation is an essential prerequisite for the legitimate authority of action-guiding legal norms.").

83. *See* LAURENCE H. TRIBE, AMERICAN CONSTITUTIONAL LAW 666 (2d ed. 1988) (distinguishing between the "instrumental" and "intrinsic" values of due process).

84. *See* Selbst & Barocas, *supra* note 8, at 1120–22. This strategic behavior is sometimes called "gaming" the algorithm, but that is usually reserved for treating the behavior negatively. *See* Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 1 (2018); Ignacio N. Cofone & Katherine J. Strandburg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. 623, 623 (2018).

85. *See* Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 2003 (2021) ("Reason giving is central to contestation . . . ."); Solon Barocas, Andrew D. Selbst & Manish Raghavan, *The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 80, 80–82 (2020); Suresh Venkatasubramanian & Mark Alfano, *The Philosophical Basis of Algorithmic Recourse*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 284, 285 (2020).

86. *See* Kaminski & Urban, *supra* note 85, at 2003; *see also* Selbst & Powles, *supra* note 76, at 236; Isak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET L. 77, 80–81 (Tatiani-Eleni Synodinou et al. eds., 2017); Wachter et al., *supra* note 8, at 872–78; *see generally* Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond*, 27 INT'L J.L. & INFO. TECH 91 (2019). For more on the debate about the right to explanation, *see* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 189 (2019).

lator or lawmaker will want to know if the model comports with society's values as expressed in law. So far, the bulk of the technical work on interpretability or explainable AI is geared toward allowing engineers to troubleshoot their own products for accuracy.[87] While these are the most likely explanatory processes to be adopted, as they align with product development, they are not the kinds of explanations that justify decisions or provide recourse. But it is not clear that any mechanically generated explanation can provide justifications or recourse effectively, and certainly none can do it universally. When people explain outcomes to each other, as might be the case when speaking to a human bureaucrat, there is a discursive back-and-forth — a mutual sharing of expectations about what the explanation is attempting to achieve — that shapes the form of the explanation itself. It's a conversation. That cannot happen when a machine is programmed to "explain itself."

Just like in the discrimination case, the type of explanation that is deemed appropriate depends on a host of subjective decisions. Take as an example an explanation that highlights the most "important" factor in a decision. If the goal is to enable a consumer to get credit in the future, the explanation should highlight the factor easiest to change, but if the goal is procedural justice, the factor to highlight might be the one that makes up the bulk of the result, which might be different.[88] The goals of the explanation — here, each reasonable but incompatible — will determine which is given. Even more challenging, a given explanation of an attribute will not always correspond to obvious actions that a consumer can take. (For example, is it easier to raise one's income or stay put and increase one's job tenure? It depends.[89]) All of this is complicated further if one accounts for the possibility of a firm using the explanation to nudge consumer action in ways that benefit itself.[90]

Therefore, even where a right to explanation exists, such a right can likely be satisfied in any number of ways, which may or may not satisfy anyone's material concerns about having life opportunities denied by an inscrutable black box. Like in the discrimination case, then, the subjective rationales for the chosen approach to explanation need to be surfaced in order to contextualize the approach. How do explainers normalize the variables in their models? How do they decide what types of explanations to offer and to whom? What explana-

87. Lisa Käde & Stephanie von Maltzan, *Towards A Demystification of the Black Box-Explainable AI and Legal Ramifications*, 23 J. INTERNET L. 3, 4 (2019) ("While explainability is an essential property for decisionmaking processes, it nonetheless serves only as a useful debugging tool to detect biases in machine learning models.").

88. Barocas et al*., supra* note 85, at 83.

89. *Id.*

90. *Id.* at 86–87.

tory considerations enter into the initial model design stage? Only with answers to those questions can consumers understand the explanations they are given, and can regulators appreciate whether they are adequate. Again, these are the types of questions answered in an AIA.

C. The Unsafe Medical AI

The last example is the unsafe medical diagnostic tool. Like discrimination and opacity, safety is another overriding and oft-discussed concern within the AI accountability discourse.[91] Safety concerns are relevant any time AI interacts with the real physical world, notably including autonomous vehicles and medical devices. Medical AI is advancing rapidly, used in both diagnosis and treatment.[92] AI is particularly promising when it comes to early detection of certain maladies, like cancer[93] and sepsis,[94] for which earlier detection can mean the difference between life and death. In such a physically sensitive context, there will of course be important safety issues to address, which can stem from both errors in operation and unsecure systems.

Medical AI presents both operational and security challenges. If a cancer screening works incorrectly, it risks misdiagnosing a tumor, leading to either an absence of necessary treatment or application of unnecessary and possibly harmful treatment. If it is unsecure, it risks being manipulated from the outside to work incorrectly, ending up with the same result. Research has demonstrated that just this sort of

---

91. *See, e.g.*, Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety* (Jun. 21, 2016) (unpublished manuscript), https://arxiv.org/abs/1606.06565 [https://perma.cc/7ZB7-3MYK]; ION STOICA ET AL., A BERKELEY VIEW OF SYSTEMS CHALLENGES FOR AI 1 (2017) (discussing security challenges). "AI Safety" sometimes refers to the threat posed by a futuristic superintelligence. *See* NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 5 (2014). If that is at all realistic, it is essentially unrelated to the machine learning systems we typically refer to as AI today and is not a near-term concern. *See, e.g.*, AI NOW, THE AI NOW REPORT: THE SOCIAL AND ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE NEAR-TERM 18 (2016).

92. W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 420 (2015).

93. Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis & Dimitrios I. Fotiadis, *Machine Learning Applications in Cancer Prognosis and Prediction*, 13 COMPUTATIONAL & STRUCTURAL BIOTECHNOLOGY J. 8, 12–16 (2015) (surveying success rates of machine learning applications in cancer treatment); H.A. Haenssle et al., *Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists*, 29 ANNALS ONCOLOGY 1836, 1838–39 (2018).

94. Mark Sendak et al., *"The Human Body Is a Black Box": Supporting Clinical Decision-Making with Deep Learning*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 99, 106 (2020); Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon & A. Aldo Faisal, *The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care*, 24 NATURE MED. 1716, 1716–20 (2018) (discussing use of AI to suggest treatment options for patients diagnosed with sepsis).

manipulation is possible: AI trained to determine whether a tumor is malignant can be tricked into switching its answer.[95]

In the current legal environment, these kinds of physically harmful outcomes from AI would be addressed by whatever regulatory apparatus applies to the subject matter. Often this will be tort law — negligence if the error is attributable to improper use of the AI tool, or products liability if the error is alleged to stem from a product defect. The problem is that AI is not a comfortable fit for these tort regimes, for a number of reasons.[96] Looking at negligence, we see a similar problem to the discrimination context, based on the need to find fault. Imagine that there is a type of cancer that a reasonable doctor would always catch, but is still missed 20 percent of the time. Twenty percent of patients would have a malpractice suit, and thus some remedy. Now imagine an AI capable of reducing that error rate slightly — say to 15 percent — but as a black box, such that it would be difficult for a doctor to know when it is making an error. No hospital or doctor could be found negligent for using the AI — it would save lives! But that means that 15 percent of people would now be injured with no remedy.[97]

The way around this would be to have an AI that comes with documentation describing the populations or situations where it is less likely to work. Such documentation would instruct the medical staff about the circumstances in which they should or should not rely on the AI, or should be more willing to second-guess it. These could be for certain subpopulations that the model has a higher error rate for, or detection in the face of certain other medical conditions that make the reading more error-prone, or anything else. With documentation, the medical AI becomes more like any old medical equipment than an all-seeing oracle. But today, there is no documentation required. And as

---

95. *See* Anton S. Becker et al., *Injecting and Removing Suspicious Features in Breast Imaging with CycleGAN: A Pilot Study of Automated Adversarial Attacks Using Neural Networks on Small Images*, 120 EUR. J. RADIOLOGY 1, 4–5 (2019); Yisroel Mirsky, Tom Mahler, Ilan Shelef & Yuval Elovici, *CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning*, PROC. 28TH USENIX SEC. SYM. 461, 462–63 (2019); Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam & Isaac S. Kohane, *Adversarial Attacks on Medical Machine Learning*, 363 SCIENCE 1287, 1288 (2019) [hereinafter Finlayson et al., *Adversarial Attacks*]; *see also* Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane & Andrew L. Beam, *Adversarial Attacks Against Medical Deep Learning Systems* 2, 2–4 (2019) (unpublished manuscript), https://arxiv.org/abs/1804.05296 [https://perma.cc/VUM4-Z9PL]. Possible, of course, does not mean likely. Catherine Olsson, *Unsolved Research Problems vs. Real-world Threat Models*, MEDIUM (Mar. 26, 2019), https://medium.com/@catherio/unsolved-research-problems-vs-real-world-threat-models-e270e256bc9e [https://perma.cc/9ZQY-YXA3].

96. *See generally* Selbst, *supra* note 8 (arguing that negligence liability will be difficult to establish for users of algorithmic systems due partly to the systems' opacity and lack of documentation).

97. *Id.* at 1376.

long as the AI saves more lives than not using it, the choice to use it, even without documentation, will likely not be considered negligent.

Though purportedly strict rather than fault-based,[98] relying on products liability is not much better. Products liability hinges entirely on the question of whether a defect is present. But AI recommendations are statistical in nature. AI efficacy is judged by error rates, rather than errors in individual cases. Thus, if an AI gets 90 percent of cases right, the 10 percent of cases it gets wrong are not evidence of a defect — they are in fact the system working properly. A plaintiff who wanted to make out a products liability claim would have to argue that he was not in the 10 percent that were "supposed" to be in error, a claim that inherently makes no sense, because if the "correct" 10 percent were knowable in advance, the algorithm would fix its answers. While testing is expected in the products liability context to determine whether a product is unreasonably unsafe,[99] that would typically go to the overall error rate, and if the product saved lives, it would be hard to argue that it is unsafe overall. Once again, this could potentially be addressed with testing and documentation to understand whether error rates in a given model are not randomly distributed, but pertain to certain groups of people, sorted by demographics or medical profile. All this information would be included in an AIA, potentially enabling tort remedies to work better.[100]

In the case of medical devices specifically — as opposed to other devices that pose physical harm like cars — the Food and Drug Administration ("FDA") does have a role for ex ante examination and pre-market approval.[101] In January 2021, the FDA released an action plan to update its medical device protocols for AI,[102] responding to its earlier realization that its "traditional paradigm of medical device regulation was not designed for adaptive artificial intelligence and ma-

98. Products liability is today acknowledged to be more negligence-like than the doctrine or phrase "strict products liability" suggests. *See* William C. Powers, Jr., *The Persistence of Fault in Products Liability*, 61 TEX. L. REV. 777, 777–78, 813 (1983). This is because defects incorporate an idea of reasonableness, like negligence. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. §§ 1–2 (AM. L. INST. 1997). But that is less important for this discussion than the fact that what is judged is the product, not the conduct of any person.

99. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. §§ 1–2 (AM. L. INST. 1997).

100. There is also an argument that a failure to warn of known or reasonably knowable risks to a certain subset of potential patients would constitute a warning defect. This would suggest that products liability doctrine could already require a similar type of investigation to an AIA, though the specific contours of the two would not necessarily be an exact overlap.

101. *See* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 83 (2017) (arguing that FDA premarket approval would be a good regulatory model for algorithmic governance).

102. FOOD & DRUG ADMIN., ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (AI/ML) SOFTWARE AS A MEDICAL DEVICE ACTION PLAN (2021).

chine learning technologies."[103] Thus, though unsafe medical AI is a useful demonstrative example, medical AI may in reality turn out to be one type of technology that does have an ex ante requirement to undergo testing before launch. That said, it is likely an AIA regime would turn out to be complementary, or in fact, could be modeled in part on the FDA's processes.[104]

III. ALGORITHMIC IMPACT ASSESSMENTS

In the last five years, scholars, think tanks, policy advocates, and legislators have proposed or implemented various AIAs, or argued that existing law imposes an equivalent requirement.[105] But despite the unifying term, these AIA proposals actually differ significantly.[106] They include proposals modeled on NEPA directly,[107] arguments that the data protection impact assessments ("DPIA") of the GDPR accomplish the task,[108] and the Canadian AIA, which is a quick, effi-

---

103. *Artificial Intelligence and Machine Learning in Software as a Medical Device*, FOOD & DRUG ADMIN. (Sept. 22, 2021), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [https://perma.cc/C4U8-WHYR].

104. In fact, one of the courses of action that the FDA has specially identified is harmonization of good machine learning practice, FOOD & DRUG ADMIN., *supra* note 102, at 3–4, which is one of the benefits of an AIA mandate as well.

105. *See, e.g.*, Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019); Data Broker Accountability and Transparency Act of 2020, H.R. 6675, 116th Cong. (2020); Selbst, *supra* note 37, at 169; DILLON REISMAN, JASON SCHULTZ, KATE CRAWFORD & MEREDITH WHITTAKER, ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY 11–20 (2018); Selbst & Barocas, *supra* note 8, at 1130, 1134–35; Margot E. Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-layered Explanations*, 11 INT'L DATA PRIVACY L. 125, 125 (2020); Alessandro Mantelero, *AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment*, 34 COMPUTER L. & SEC. REV. 754, 754 (2018); Simon Reader, *Data Protection Impact Assessments and AI*, INFO. COMM'R'S OFF. (Oct. 23, 2019), https://ico.org.uk/about-the-ico/news-and-events/ai-blog-data-protection-impact-assessments-and-ai/ [https://perma.cc/U6FL-JCCC]; CANADIAN AIA, *supra* note 32; Sonia K. Katyal, *supra* note 48, at 55, 61–62 (2019).

106. *See* EMANUEL MOSS, ELIZABETH ANNE WATKINS, RANJIT SINGH, MADELEINE CLARE ELISH & JACOB METCALF, ASSEMBLING ACCOUNTABILITY: ALGORITHMIC IMPACT ASSESSMENT FOR THE PUBLIC INTEREST 29–30 (2021) ("There are already multiple proposals and existing regulations that make use of the term 'algorithmic impact assessment.' While all have merits, none share any consensus about how to arrange the constitutive components of AIAs.").

107. Selbst, *supra* note 37, at 168; REISMAN ET AL., *supra* note 105, at 7; EUROPEAN PARLIAMENTARY RESEARCH SERVICE, A GOVERNANCE FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY AND TRANSPARENCY 52–63 (2019). These proposals, in turn, built on prior instances of NEPA-derived models in the information law space, such as the Privacy Impact Assessment (PIA) mandated by the E-Government Act of 2002, and Michael Froomkin's proposal of a "Privacy Impact Notice." *See* A. Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements*, 2015 U. ILL. L. REV. 1713, 1745–55 (2015).

108. Kaminski & Malgieri, *supra* note 105, at 129–33.

cient questionnaire.[109] These proposals also exist alongside similar proposals to implement a regime of algorithmic auditing,[110] a regulatory cousin of AIAs, as well as industry self-regulation that seeks to implement AIAs and audits as a matter of ethics.[111]

To the extent it is a distinct approach from other approaches to regulation and oversight, an impact assessment approach should have characteristics dictated by its twin goals. The design goal aims to get project developers to use their expertise to estimate future impacts of the various decisions before implementation, so that any anticipated harmful impacts can be eliminated or mitigated early, rather than corrected after the fact, when it can be costlier or outright impossible to do so. The knowledge production goal seeks to use and export the information generated — the estimation of future impacts, as well as documentation about the decision process — for accountability purposes or the development of future policy. Though, as Part IV will discuss, these two goals can be tricky to achieve together in practice, keeping them in focus has certain implications for the design of AIA regulation. Not all existing proposals effectuate these goals well, so this Part examines the various proposals to see how and if each implements the core aspects of the AIA model that make it useful.

A. The AIA Models

The AIA proposals essentially fit into three categories: (1) models based on NEPA, (2) models based on the GDPR's DPIA, and (3) a questionnaire model. Self-regulatory or "ethics" models of impact assessment and audits each present other options. I will briefly describe them all here before turning to the differences.

---

109. CANADIAN AIA, *supra* note 32 (presenting an AIA consisting of "around 60 questions related to [] business process, data and system designed decisions").

110. *See* ADA LOVELACE INST., EXAMINING THE BLACK BOX: TOOLS FOR ASSESSING ALGORITHMIC SYSTEMS 7 (2020) ("There are two methodologies that have seen wide reference in popular, academic, policy and industry discourse around the use of data and algorithms in decision making: algorithm audit and algorithmic impact assessment."); Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143, 152 (2019) (arguing that "data auditing methodologies . . . will likely become the new norm for promoting compliance in automated systems"); Christian Sandvig, Kevin Hamilton, Karrie Karahalios & Cedric Langbort, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, 64 ANN. MEETING INT'L COMMC'N ASS'N 1, 8 (2014), http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf [https://perma.cc/QBB6-9WKN]; CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION 208 (2014).

111. *E.g.*, Jovana Jankovic, *U of T's Schwartz Reisman Institute and AI Global to Develop Global Certification Mark for Trustworthy AI,* U T NEWS (Dec. 1, 2020) https://www.utoronto.ca/news/u-t-s-schwartz-reisman-institute-and-ai-global-develop-global-certification-mark-trustworthy-ai [https://perma.cc/Z6ZD-W2LU].

Proposals by the think tank AI Now and the European Parliamentary Research Service, as well as an earlier proposal of mine, are all NEPA-based models.[112] The NEPA model implies a highly detailed impact assessment, potentially running to hundreds of pages.[113] It demands thorough answers to open-ended questions that explain the design process.[114] Other features of the NEPA model are transparency and public participation via a notice and comment framework.[115] Because transparency, and specifically notice and comment frameworks, are part of the regulation that is usually applied to the public sector in the United States, it is perhaps not surprising that these proposals tend to focus on the public sector, rather than the private sector.

The second model for AIAs draws on European data protection law. Article 35 of the GDPR requires companies to perform DPIAs whenever data processing "is likely to result in a high risk to the rights and freedoms of natural persons."[116] Margot Kaminski and Gianclau-

---

112. Selbst, *supra* note 37, at 168; REISMAN ET AL., *supra* note 105, at 4; EUROPEAN PARLIAMENTARY RESEARCH SERVICE, A GOVERNANCE FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY AND TRANSPARENCY 52–63 (2019). These proposals, in turn, built on prior instances of NEPA-derived models in the information law space, such as the Privacy Impact Assessment ("PIA") mandated by the E-Government Act of 2002, and Michael Froomkin's proposal of a "Privacy Impact Notice." *See* A. Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements*, 2015 U. ILL. L. REV. 1713, 1755 (2015).

113. *See* 40 C.F.R. § 1502.7 (2020) (limiting an EIS to 300 pages in extraordinary circumstances). EISs have been known to exceed that limitation, *see* Michael Herz, *Parallel Universes: NEPA Lessons for the New Property*, 93 COLUM. L. REV. 1668, 1713 (1993) (discussing "[t]he sprawling, unfocused, thousand-page EIS with twenty-eight appendices"), and the focus on information quantity over quality is a frequently cited complaint with NEPA as a framework. *See, e.g.*, Alyson C. Flournoy, Heather Halter & Christina Storz, *Harnessing the Power of Information to Protect Our Public Natural Resource Legacy*, 86 TEX. L. REV. 1575, 1582–83 (2008).

114. To be sure, not every instance of an impact assessment under this model will be so detailed in reality, but those cases can be considered not substantially in compliance with the requirement. *Cf.* Kenneth A. Bamberger & Deirdre K. Mulligan, *PIA Requirements and Privacy Decision-Making in US Government Agencies*, *in* 6 PRIVACY IMPACT ASSESSMENT 225, 225–50 (D. Wright & P. De Hert eds., 2012) (comparing the two federal agencies' vastly different PIAs regarding the use of RFID chips).

115. *See* 40 C.F.R. § 1503.1 (2020) (describing the notice and comment process for NEPA's EIS); Selbst, *supra* note 37, at 177 (applying the same to the AIA framework); REISMAN ET AL., *supra* note 105, at 7–11 (describing a notice and comment procedure for their proposal).

116. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, art. 35(1), 2016 O.J. (L 119) 53. On its face, this requirement appears to narrow the scope of situations in which a DPIA would be required as compared to the NEPA model, but that is not obvious upon closer inspection. The way that "high risk" is interpreted in official interpretations by the Article 29 Working Party implies that the requirement will apply in a broad array of cases involving data-driven technologies. *See* Working Party on the Protection of Personal Data 95/46/EC, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679, art. 29, WP 248 (Apr. 4, 2017) [hereinafter DPIA Guidance]; *see also* Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking for*, 16 DUKE L. & TECH. REV. 18, 78 n.243 (2017) (citing DPIA Guidance, *supra*) ("Judging by this guidance,

dio Malgieri have argued that when read against the broader background of the interwoven rights and oversight provisions in the GDPR, this requirement should be read to encompass an AIA.[117] Some EU member state governments also read the GDPR as requiring a version of an AIA. As Kaminski and Malgieri note, Slovenia has already read the GDPR's explanation requirements under Article 22 to include an AIA.[118] And before the United Kingdom left the EU, the UK Information Commissioner's Office ("ICO") released a draft of their AI Auditing Framework for comment.[119] In it, the ICO stated that the GDPR's "accountability principle" requires a DPIA for all automated processing.[120]

The DPIA envisions a similarly expansive scope of work to the NEPA model, including a "systematic description" of the processing, justifications, and plans for mitigation.[121] One difference from the NEPA approach is that there is no explicit requirement to describe all the reasonable and rejected choices. The only requirement is to systematically evaluate the actual program that is to go forward. In practice, however, the requirement to show all the "measures envisaged" to mitigate dangers might be broad enough to encompass the same idea.[122] The most significant difference is in transparency. Although the official guidance on DPIAs recommends making a summary of the DPIA public, publication — of even a summary — is not required.[123] Instead DPIAs are performed in collaboration with member

almost every ML system seems likely to require a DPIA."); Casey et al., *supra* note 110, at 173 ("The A29WP's guidance stresses that, in many circumstances, DPIAs are not merely recommended as a matter of best practices but are compulsory."). As Bryan Casey and colleagues note, "demonstrating that a DPIA is not necessary will, in many instances, itself require a DPIA." *Id.* at 175.

117. Kaminski & Malgieri, *supra* note 105, at 129–33; *see also* Yordanka Ivanova, *The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI*, PRIV. TECH. & POL'Y 1, 3 (2020); Heleen L. Janssen, *Detecting New Approaches for a Fundamental Rights Impact Assessment to Automated Decision-Making*, 10 INT'L DATA PRIV. L. 76, 88 (2020) (arguing that DPIAs "will become one of the mechanisms for the governance of fundamental rights in [automated decision-making]"). *But see* Mantelero, *supra* note 105, at 756, 761–62 (arguing that Article 35 of the GDPR does not adequately address the ethical and social issues presented by AI).

118. Kaminski & Malgieri, *supra* note 105, at 129.

119. INFO. COMM'R'S OFF., GUIDANCE ON THE AI AUDITING FRAMEWORK 12 (2020). The "accountability principle" is an affirmative duty to ensure and be able to demonstrate compliance with the substantive provisions of the GDPR, and it is codified in Article 5(2): "The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability')." Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, art. 5(2), 2016 O.J. (L 119) 36.

120. INFO. COMM'R'S OFF., *supra* note 119.

121. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, art. 35(7), 2016 O.J. (L 119) 54.

122. *Id.*

123. *See* DPIA Guidance, *supra* note 116, at 18.

state data protection authorities, as is more typical of the more collaborative European style of data protection regulation.[124]

The third approach is the one taken by the government of Canada. Under Canada's Directive on Automated Decision-Making,[125] government agencies that use algorithmic decisionmaking must complete an AIA both before production and before the project goes live. The AIA consists of "around 60 questions related to [] business process, data and system designed decisions."[126]

The questions touch on most of the topics people care about with respect to algorithms. Some of the questions go to the thoughts behind the process (e.g., "What is motivating your team to introduce automation into this decision-making process? (Check all that apply),"[127] with choices related to backlog, efficiency, quality, and being innovative).[128] Other questions ask about the stakes of the decisions, the sector, the degree of explanation or human involvement, and so on.[129] Each of these questions receive a point total. That point total then determines whether the overall risk falls within one of four wide bands (Impact Levels I–IV), and agencies implementing algorithmic system that fall within a given band must take certain increasingly involved remedial actions to mitigate the anticipated harms.[130] While most of the questions are multiple choice, some do include written answers.[131] The written answers are not scored, but can be made public.[132]

These approaches can be additionally contrasted with the push for better documentation and impact assessments from within the technology industry. In general, as part of the corporate social responsibility movement in the last half-century, companies have developed methodologies for "social impact assessment" ("SIA")[133] even in the

---

124. *See* Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CALIF. L. REV. 1529, 1530 (2019); William McGeveran, *Friending the Privacy Regulators*, 58 ARIZ. L. REV. 959, 965–979 (2016).

125. GOVERNMENT OF CANADA, TREASURY BOARD, Directive on Automated Decision-Making (2019), § 6.1, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [https://perma.cc/JH5R-AYTY].

126. CANADIAN AIA, *supra* note 32.

127. *Id*.

128. *Id.*

129. *Id.*

130. GOVERNMENT OF CANADA, TREASURY BOARD, Directive on Automated Decision-Making (2019), Appendix C, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [https://perma.cc/JH5R-AYTY].

131. *See* CANADIAN AIA, *supra* note 32.

132. *See* GOVERNMENT OF CANADA, TREASURY BOARD, Directive on Automated Decision-Making (2019), §§ 4.2.3, 6.1.4, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [https://perma.cc/JH5R-AYTY].

133. *See generally* Freudenburg, *supra* note 35; Esteves et al., *supra* note 35. There is actually an association of impact assessment professionals known as the International Association for Impact Assessment ("IAIA") that publishes its own academic journal. *Impact Assessment and Project Appraisal*, INT'L ASS'N IMPACT ASSESSMENT https://www.iaia.org/iapa-journal.php [https://perma.cc/6UGM-S5ZU]. The field is broadly

absence of regulatory requirements.[134] The earliest public call for impact assessments in the algorithmic context was in this vein. In 2016, a group of scholars who work on AI fairness published a document titled "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms."[135] The document laid out five high level principles — responsibility, explainability, accuracy, auditability, and fairness.[136] It also outlined what an SIA for algorithms could look like, recommending a set of probing questions that should be answered three times: at the design stage, pre-launch, and post-launch.[137] This document works well as a guide to the principles and practice of an SIA for algorithms. Since then, the push for impact assessments has been taken up as part of the broader call for responsible AI or "AI ethics."[138]

A similar self-regulatory push comes in the form of human rights impact assessments ("HRIA"). A framework for conducting HRIAs is recommended by the UN Guiding Principles on Business and Human Rights,[139] and companies who are conducting them may do so because they feel a moral obligation to respect human rights or because their social license demands it,[140] as likely was the case after Facebook's negligence assisted genocide in Myanmar.[141] But there is no

---

and highly interdisciplinary and appears to be a great resource for detailed methods and methodologies for performing impact assessments.

134. *See* Esteves et al., *supra* note 35, at 35–36 (noting that SIA is commonly used for project approvals but has expanded beyond regulatory requirements).

135. Diakolopous et al., *supra* note 35. Sonia Katyal has additionally proposed a Human Impact Statement based on the NEPA model, but did not discuss mandates, suggesting that it is appropriately seen as a self-regulatory proposal. *See* Katyal, *supra* note 48, at 115–17.

136. *Id.*

137. *Id.*

138. *See, e.g.,* Rafael A. Calvo, Dorian Peters & Stephen Cave, *Advancing Impact Assessment for Intelligent Systems*, 2 NATURE MACH. INTEL. 89, 89 (2020); Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti & Michael Lennon, *Principles to Practices for Responsible AI: Closing the Gap* (June 8, 2020) (unpublished manuscript) (on file at https://arxiv.org/abs/2006.04707 [https://perma.cc/98VN-3S69]) (noting that "[a]s of 2019, more than 20 firms have produced frameworks, principles, guidelines, and policies related to the responsible development and use of artificial intelligence (AI)," and arguing that "an impact assessment framework which is broad, operationalizable, flexible, iterative, guided, and participatory is a promising approach to close the principles-to-practices gap"); *cf.* Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 33, 38 (2020) (calling for an audit practice to "increase ethical foresight").

139. *See* U.N. Human Rights Office of the High Commissioner, Implementing the United Nations "Protect, Respect and Remedy" Framework, U.N. Doc. HR/PUB/11/04 iv (2011).

140. Sara Bice & Kieren Moffat, *Social License to Operate and Impact Assessment*, 32 IMPACT ASSESSMENT & PROJECT APPRAISAL 257, 257 (2014).

141. *See* Paul Mozur, *A Genocide Incited on Facebook, With Posts from Myanmar's Military*, N.Y. TIMES (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html [https://perma.cc/MQ3C-ME62] (discussing Facebook's role in the genocide in Myanmar); MARK LATONERO, GOVERNING ARTIFICIAL INTELLIGENCE: UPHOLDING HUMAN RIGHTS & DIGNITY 18–19 (2019),

particular human rights law requirement to conduct HRIAs, and the most prominent example — Facebook's post hoc HRIA for its role in Myanmar[142] — was considered a failure that acted more like "ethics washing" than anything substantive.[143]

Both SIAs and HRIAs can be useful tools. The assessments themselves may be substantively equivalent to AIAs that would be required by a law. But it will not be enough to rely on self-regulation.[144] The movement for AI Ethics has grown from a few statements of ethical principles and frameworks into an entire self-regulatory compliance industry in a few short years.[145] But perhaps predictably, that industry is failing to address the social issues that AIA regulation sets out to solve.[146]

Finally, industry actors and scholars have been proposing algorithm audits separately from impact assessments.[147] Some scholars

---

https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf [https://perma.cc/T78W-YE64] (describing the public pressure on companies to perform HRIAs, including Facebook's HRIA about its involvement Myanmar).

142. BSR, HUMAN RIGHTS IMPACT ASSESSMENT: FACEBOOK IN MYANMAR (2018) https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf [https://perma.cc/MSJ9-PLJW].

143. *See* MARK LATONERO & AAINA AGRAWAL, HUMAN RIGHTS IMPACT ASSESSMENTS FOR AI: LEARNING FROM FACEBOOK'S FAILURE IN MYANMAR 1–2 (2021), https://carrcenter.hks.harvard.edu/publications/human-rights-impact-assessments-ai-learning-facebook%E2%80%99s-failure-myanmar [https://perma.cc/GU9C-KEC2]; JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM 241 (2019) ("Guiding principles and special reports intended to constrain corporate conduct have no independent legal force, [] and the unprecedented power of capital over the conditions of human freedom has continued to grow."); *id.* at 245 (discussing how human rights in the corporate world have been brought under the umbrella of corporate social responsibility).

144. Self-regulation could, in turn, be bolstered by an agency such as the Federal Trade Commission that can hold industry to its promises, rendering it what some scholars call "soft law." *See* Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournaso, *Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance*, 61 JURIMETRICS 133, 137 (2020). For reasons explained in Section III.B.1, *infra*, such a solution is likely too easily captured by industry to be useful.

145. *See* Anna Jobin, Marcello Ienca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 NATURE MACH. INTEL. 189, 389 (2019); Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NATURE MACH. INTEL. 501, 501 (2019) ("To date, at least 84 'AI Ethics' initiatives have published reports describing high-level ethical principles, tenets, values, or other abstract requirements for AI development and deployment."); Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, 30 MINDS & MACHS. 99, 99–100 (2020).

146. *E.g.*, Karen Hao, *In 2020, Let's Stop AI Ethics-Washing and Actually Do Something*, MIT TECH. REV. (Dec. 27, 2019), https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/ [https://perma.cc/GEN4-6PBM].

147. *See* ADA LOVELACE INST., EXAMINING THE BLACK BOX: TOOLS FOR ASSESSING ALGORITHMIC SYSTEMS 7 (2020) ("There are two methodologies that have seen wide reference in popular, academic, policy and industry discourse around the use of data and algorithms in decision making: algorithm audit and algorithmic impact assessment."); Casey et al., *supra* note 110, at 152 (arguing that "data auditing methodologies . . . will likely become

have argued that only audits, and not impact assessments, can rein in companies.[148] Audits are similar in concept to impact assessments. Indeed, the two ideas sometimes bleed together enough that distinguishing them in general is difficult.[149] Sometimes audits are seen as checking off best practice checklists, sometimes they are tests of particular input-output relationships — such as housing audits common in the fair housing context — and sometimes they are involved, back-and-forth processes with regulators intensely monitoring the practices of companies, such as regulatory audits in the finance context. As a result of all this language confusion, recommendations for audits as a regulatory tool end up meaning many different things,[150] and now an industry is forming that claims to do audits for a fee, without any sort of standardization.[151] So yes, audits are likely useful, but audits and impact assessments overlap significantly as concepts, and it is not clear that either is so rigidly defined a concept that one can be said to solve anything the other does not. Rather than making further definitional distinctions that will inevitably be contested, I lay out the functional aspects of what I think makes an AIA useful below. Any audit that operates substantially the same way can be included in that grouping, regardless of label.

B. The Important Aspects of an AIA

The different AIA models can all potentially be useful, but it is important when structuring an AIA requirement to foreground the two goals: (1) changing design processes to consider social harms in early stages, and (2) providing documentation to enable accountability and policy-learning. The NEPA and DPIA models are both potentially compatible with these goals, while the Canadian AIA is less so. This Section explains why.

1. Early Intervention

As an example of reflexive regulation,[152] impact assessment frameworks are meant to be early-stage interventions, to inform pro-

---

the new norm for promoting compliance in automated systems"); Sandvig et al., *supra* note 110, at 5; O'NEIL, *supra* note 110, at 208.

148. Ari Ezra Waldman, *Privacy Law's False Promise*, 97 WASH. U. L. REV. 773, 806–07 (2020) (discussing securities audits as a model).

149. *See generally* ADA LOVELACE INST., *supra* note 147, at 3–4, 8, 15.

150. *See generally* RYAN CARRIER & SHEA BROWN, TAXONOMY: AI AUDIT, ASSURANCE & ASSESSMENT (2021), https://forhumanity.center/blog/taxonomy-ai-audit-assurance-assessment [https://perma.cc/LEU8-JUYF].

151. *See* Ng, *supra* note 38.

152. *See infra* Section III.A.

jects before they are built.[153] This timing restriction is inherent to the first goal. Perhaps less obviously, the timing restriction is also important for the second goal of enabling the public to learn from the design choices. In complex enough systems, post hoc explanation is challenging, if not impossible. Many design factors interact in hidden ways, such that trying to parse a causal explanation for some error is impossible, unless we have documentation about what choices were made and why, and we can trace out a counterfactual world of decisions that might otherwise have been made. This is certainly true for algorithmic systems, where the inherent complexity has led to a great deal of research on how to make models interpretable.[154] Whether we're concerned about discrimination, explanation, safety, or something else, we need to know how the many subjective decisions that go into building a model led to the observed results, and why those decisions were thought justified at the time, just to have a chance at disentangling everything when something goes wrong.

The AIA proposals all work this way, so it is not a difference between them. The NEPA and DPIA models require completion of the AIA before deployment of the project.[155] The Canadian AIA is meant to be filled out before design and again after implementation.[156] The timing restriction is still worth noting, however, as SIAs, HRIAs, and audits — governance practices often lumped in with AIA — can sometimes occur entirely after the fact.[157]

---

153. There is a good analogy here to the concept of "technical debt," the idea that if early on in a project one makes certain easier decisions to get something to work, then one will have to pay back later with greater effort. *See* Henriette Cramer, Jean Garcia-Gathright, Aaron Springer & Sravana Reddy, *Assessing and Addressing Algorithmic Bias in Practice*, 25 ACM INTERACTIONS 59, 62 (2018) ("Algorithmic bias to a certain extent can be seen as technical debt. Bias is much easier to tackle when working with a new product rather than one that has been running for a while, or where a variety of models are working together. Unintended biases are self-reinforcing, recursive, and much harder to eliminate if ignored at the beginning.").

154. *See generally, e.g.*, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti & Dino Pedreschi, *A Survey of Methods for Explaining Black Box Models*, ACM COMPUTING SURVEYS, 2019, at 1; Zachary C. Lipton, *The Mythos of Model Interpretability*, PROC. ICML WORKSHOP ON HUM. INTERPRETABILITY MACH. LEARNING 96 (2016).

155. *See* 40 C.F.R. § 1502.2 (2020) ("Environmental impact statements shall serve as the means of assessing the environmental impact of proposed agency actions, rather than justifying decisions already made."); GDPR art. 35 (1) (requiring a DPIA "prior to the processing" of data).

156. CANADIAN AIA, *supra* note 32, § 3.1.

157. *See* Frank Vanclay & Philippe Hanna, *Conceptualizing Company Response to Community Protest: Principles to Achieve a Social License to Operate*, 8 LAND 101 (2019) (discussing SIAs as response to community protest in general); LATONERO & AGRAWAL, *supra* note 143, at 19 ("One of the tragedies of the Facebook Myanmar issue is that the company apparently did not respond adequately to repeated attempts by civil society, human rights groups, and academic researchers to alert the company that hate groups were using the platform to harm other users."). Audits are often post hoc but can also be ongoing pro-

2. Open-Ended Questions

An effective AIA must ask open-ended questions, inviting bottom-up explanations. The algorithmic systems of interest are highly complex and far from fully understood, with many unknown unknowns. A major benefit of an impact assessment regime is that there is a bottom-up reporting structure. Rather than ask if specific checks were completed, like an audit might, an AIA can require the designers to explain their decisions.

This is the main difference between the NEPA model or DPIAs and something like the Canadian questionnaire model. The NEPA model would instruct the assessor to, among other things, "rigorously explore and objectively evaluate all reasonable alternatives."[158] This is a flexible requirement that anticipates that any attempt to anticipate answers will necessarily leave some out. While the questions that the Canadian AIA asks are thoughtful, they are fixed and quite general. Here are a few of the questions that the Canadian AIA asks about input data (where choices are provided, they are the contents of drop-down menus, and the rest are yes/no questions):

(1) Will the Automated Decision System use personal information as input data?

(2) Who controls the data?

  (a) Open Data Source

  (b) Federal government

  (c) Other Canadian Government (prov/municipal)

  (d) Private Sector/NGO

(3) Will the system use data from multiple different sources?

(4) Who collected the data used for training the system?

  (a) Your institution

  (b) Another federal institution

  (c) Another level of government

---

cesses that take place "throughout the internal organization development lifecycle." Raji et al., *supra* note 138, at 33.

158. Selbst, *supra* note 37, at 173. The NEPA model does also try to rein in the length caused by such an expansive requirement. *See* 40 C.F.R. § 1502.2 (2020) (requiring the EIS to "be analytic rather than encyclopedic," to "be discussed in proportion to their significance," and to "be no longer than absolutely necessary to comply with" the statute and regulations).

(d) A foreign government or third-party[159]

The questions asked in the Canadian AIA are important, but limited. Are the implications of all foreign government or third-party-collected datasets the same? What are the ramifications of the choice to use personal data as inputs or not?

Compare these questions to an open-ended version of the same questions that might say: "Describe all the data sources you used to train the model, including where they came from, who controls them, why you chose to use those datasets, and what impacts you anticipate from those choices." With open-ended questions, you do not need to anticipate the particular problems that might come up, and the answers to them emerge naturally. With top-down questions, no matter how thoughtful they are, the picture will be coarse and general.

Suppose one attempted to use this model of AIA for a discriminatory hiring algorithm, asking the questions listed above. Assume that the answers are "yes" to using personal data and "no" to using data from multiple sources. What do we learn from this that can even help make hiring models less discriminatory in the future? We do not know whether the personal data used increased or decreased the disproportionate results, nor do we know anything about the single or multiple sources used. While it is of course important to know whether the model uses single or multiple sources, it's much more useful to understand why the choice was made. We could make some guesses as to reasons — generally using multiple sources is more expensive, while a single source of data is more likely to be biased — but even assuming those are the correct reasons, we do not have the information to determine whether the choice was justified. Perhaps only one dataset exists in the particular subject of interest, or other data is prohibitively expensive — either would be important to know. We just do not learn much from multiple choice questions.

The problem becomes more acute when one considers differences across sectors and contexts. Early in the questionnaire, the Canadian AIA asks what sector the project is in.[160] But the next set of questions does not change depending on the answer. As a result, the questions are quite general, as seen in the input data section. Surely, the concerns we have about whether and how personal data is used are different in the hiring context, the credit context, and the medical context. In order to have top-down questionnaires that get to the important issues within a given context, there would need to be separate ques-

159. CANADIAN AIA, *supra* note 32.

160. *Id*. Choices include: "Health related services," "Economic interests (grants and contributions, tax benefits, debt collection)," "Social assistance (e.g., disability claims, etc.)," "Access and mobility (security clearances, border crossings)," "Licensing and issuance of permits," and "Other (please specify)."

tionnaires for each context. Open-ended questions do not have this same limitation.

Closed-ended questioning also limits our viewpoint into what we consider to be the relevant harms. This limitation is colorfully illustrated in a satirical paper by Os Keyes, Jevan Hutson, and Meredith Durbin, in which they imagine a unique technical solution for food shortages — turning the elderly into food.[161] In the paper, the authors applied various algorithmic audit frameworks to verify that their proposed algorithm to decide which elderly people to "render[] down into a fine nutrient slurry" was fair, accountable, and transparent.[162] In order for the closed questions in certain types of audits to prove useful, the auditor must already know what good and bad outcomes look like, and must ask the right questions.[163] If your problem with turning the elderly into food is that you may not be choosing the *right* elderly people because the algorithm is biased, you have a scoping problem in your oversight.

One last aspect of the top-down model falls short. While it may succeed at getting designers to think about problems early in design — Canada's approach has them fill out the AIA before design to prime the designers with the questions[164] — it cannot further educate regulators or the public as to the kinds of pressures, choices, and tradeoffs that the engineers and their managers must make in real practice. We can only ever ask about the problems we already know to be problems, but if the rationales are documented and explained in an open way, and we discover a problem later, we can go back and analyze what choices may have led to the problem. If the AIA is a top-down questionnaire, it does not allow us to ask questions that we do not yet know to ask. Closed-ended questioning thus fails entirely with respect to the second goal of AIAs.

3. Accountability

The difference between self-regulatory impact assessment approaches and actual regulation is, to state the obvious, that the second imposes a legal requirement. This usually also implies some form of oversight or accountability mechanism. But the actual mechanism differs between the models. The NEPA model uses transparency and

---

161. Os Keyes, Jevan Hutson & Meredith Durbin, *A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry*, 2019 CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1, 3 (2019).

162. *Id.* at 3–7.

163. *See* ADA LOVELACE INST., *supra* note 110, at 10 ("[B]ias audits cannot give a holistic picture of the system; a bias audit showing that a system doesn't treat people differently by gender does not mean the system is free of other forms of discrimination issues, or that it might not have other issues or impacts on society to be aware of.").

164. CANADIAN AIA, *supra* note 32, § 3.1.

the need to respond to public comment as its oversight mechanism. The model would require a draft AIA, followed by a public notice and comment period, then followed by a final AIA that is responsive to comments.[165] The DPIA model, by contrast, would have companies work together with regulators to generate AIAs, but without necessarily publishing the AIAs at all.[166] Each has reasons to recommend it.

Mandates for transparency and public comment apply well to the public sector, motivated by the same due process concerns that inform oversight of government action generally. But the same transparency ideal that is typically presumed to apply to the public sector does not apply to the private sector, so it is less obvious that transparency is required or always desirable.[167] Unlike in the public sector, transparency in the private sector may implicate intellectual property rights, such as trade secrets, in a way it does not for the government. Additionally, notice and comment proceedings are burdensome and slow, an imposition that we probably tolerate a great deal more when applied to the public sector than private. Finally, as discussed more in Part IV, greater transparency will risk greater resistance from the industry actors, creating incentives for them to report more vague documentation in order to protect information from competitors and public scrutiny.

Fortunately, accountability, not transparency per se, is what matters for this legislation. As Mike Ananny and Kate Crawford have detailed, for transparency to succeed in its accountability functions, there separately needs to be some mechanism to convert the knowledge gained into meaningful corrective action, and sometimes transparency can even be harmful or misleading.[168] Thus, if a model of accountability exists that does not rely on transparency, it may make sense here. The DPIA model works in this mold, relying on responsive regulation and involvement from data protection authorities in lieu of publishing of the impact assessment. As detailed at length below, AIAs will require buy-in from the private sector, and compa-

---

165. *See supra* Section III.A.

166. *See supra* Section III.A.

167. Private sector actors can be required to participate in transparent impact assessments, but typically where there is government action involved. NEPA, for example, requires "major Federal actions" for an EIS requirement, which means that the EIS requirement can be triggered by a federal funding or the need for a permit for private action. *Md. Conservation Council, Inc. v. Gilchrist*, 808 F.2d 1039, 1042 (4th Cir. 1986). Essentially a project requires an EIS if it cannot "begin or continue without prior approval of a federal agency." *Id.* State environmental protection acts have similar requirements of government action. *See* LEXISNEXIS, *Environmental Law — Assessment & Information Access: Environmental Impact Review* (Aug. 2019), https://plus.lexis.com/api/permalink/65adf888-aa0b-4c32-8a10-c7df6f7b03f2/?context=1530671 [https://perma.cc/T74B-6X7H].

168. *See generally*, Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20(3) NEW MEDIA & SOC'Y 973, 973–89 (2018).

nies will certainly be more willing to participate in process that is more flexible, less onerous, and less public. But without transparency, the specter of regulatory capture looms even larger than usual.[169]

Thinking back to the two goals of AIA regulation, if accountability for substantively completing the AIA can be achieved, then the first goal should be accomplished. The second goal, policy-learning, would seem to require a degree of transparency, as it is all about exporting knowledge to the public. But if regulators take what they learn from companies and distill that knowledge into regular reports, the second goal can be accomplished without strict transparency.[170] The ultimate determination on the issue of accountability should therefore be based on which version is likely to get better substantive compliance with the requirement.

IV. THROUGH AN INSTITUTIONAL LENS

Now we understand the two goals of the AIA and how the proposed models would implement the core concerns of early intervention, open-ended questioning, and some measure of accountability. At this point in the discussion, the AIA regulation is an idealized hypothetical. We know how it should work if everything goes to plan. But the main premise of this Article is that this ideal of an AIA regulation will never come to pass. By its nature, any AIA regulation will require some degree of cooperation from the regulated technology firms. Whenever a regulatory regime requires collaboration with regulated entities, there are significant risks to the efficacy of the regulation. Understanding how the AIA will play out on the ground, implemented by the regulated firms, is crucial to understanding the limitations of the AIA regime and how to make it as effective as possible.

Thus, this Part looks to regulatory and organizational theory to understand how an AIA regime is likely to be affected by the institutional culture and priorities of technology firms. Drawing on different ideas from the theory — collaborative governance, legal managerialism, beyond-compliance behaviors, and institutional isomorphism — this Part makes three arguments about the AIA regimes' effectiveness. First, individual firms are likely to be able to undermine substantive compliance with the AIA requirement. Thus, the first goal of AIAs — that companies consider social effects early in the process — will not be fully achievable through mere enforcement. Instead, companies

169. Kaminski, *supra* note 124, at 1529.

170. *See, e.g.*, Arti K. Rai, Isha Sharma & Christina Silcox, *Accountability, Secrecy, and Innovation in AI-Enabled Clinical Decision Software*, 7 J.L. & BIOSCIENCES 1, 23–25 (2020) (arguing that in the clinical context, the FDA should make summary information on AI development available and can do so without threatening companies' competitive interests).

will have to have internal reasons to operate in good faith. Second, the policy-learning goals of AIAs can succeed despite this challenge because only partial compliance is required. Third, despite the difficulty with enforcement, firms are not monolithic; ethical industry leaders can be used to pull the rest of the industry along behind them in the long run.

A. Collaborative Governance

At the time of its creation, the impact assessment was an innovative approach to regulation. Before NEPA, environmental decisionmaking was ad hoc, with impacts not routinely considered, alternative projects not explored, no transparency, and poor coordination between governing agencies.[171] With the country facing ever more complex environmental challenges, NEPA aimed to inject a "systematic, interdisciplinary approach" into decisionmaking about the environment, requiring government agencies to integrate science and design principles at the planning stage.[172] Rather than seeing the problem as one where political authorities needed to directly control outcomes, Congress saw the environmental challenge as too complex and broad for that. The solution was to change the way we process and account for information in making decisions.

Since then, such approaches have become common across many different contexts and governments in both the public and private sectors. Over the last two-plus decades, regulatory theorists have largely embraced collaborative governance, a suite of hybrid private-public governance models that aim to chart a course between top-down, state-centered approaches on the one side and total deregulation on the other.[173] Collaborative governance is a response to the complexity

---

171. *See* SERGE TAYLOR, MAKING BUREAUCRACIES THINK 11–13 (1984).

172. 42 U.S.C. § 4332.

173. *See* Orly Lobel, *The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought*, 89 MINN. L. REV. 342, 443 (2004); Michael P. Vandenbergh, *The Private Life of Public Law*, 105 COLUM. L. REV. 2029, 2037 (2005); Nina A. Mendelson, *Private Control over Access to the Law: The Perplexing Federal Regulatory Use of Private Standards*, 112 MICH. L. REV. 737, 747 (2014) (defining "collaborative governance" as "the public enlisting of private institutions and resources in the process of governance"); *see generally* Chris Ansell, *Collaborative Governance*, *in* THE OXFORD HANDBOOK OF GOVERNANCE 498, 498–509 (David Levi-Faur, ed., 2012). I use the phrase "collaborative governance" to refer to the umbrella of hybrid approaches that is sometimes instead referred to as "new governance." *See* Orly Lobel, *New Governance as Regulatory Governance*, *in* THE OXFORD HANDBOOK OF GOVERNANCE, *supra*, at 65, 65–79. Some scholars using the new governance terminology will refer to collaborative governance as a particular strategy under the umbrella of new governance. *See, e.g.*, Lisa Blomgren Bingham, *The Next Generation of Administrative Law: Building the Legal Infrastructure for Collaborative Governance*, 2010 WIS. L. REV. 297, 300 n.7 (2010) ("A number of legal scholars have recently examined the new governance, which includes collaborative governance, in a variety of contexts."). Overall, there does not seem to be consensus about the

of modern society and the perceived failures of state-centric administration: its inflexibility,[174] its reliance on centralization and hierarchy,[175] a lack of relevant expertise,[176] and coarsely tailored uniform rules that often miss the mark.[177] But while collaborative governance is fundamentally focused on bringing in nongovernmental perspectives, it is not simply self-regulation.[178] It is concerned with leveraging the benefits of the private sector, while keeping the accountability and legitimacy that is traditionally associated with public regulation.[179]

As a concept, collaborative governance is not crisply defined and includes many overlapping terms and models.[180] The different models have recurring themes: "(1) collaborative process, (2) stakeholder participation, (3) local experimentation, (4) public/private partnership, and (5) flexible policy formation, implementation, and monitoring."[181] Each of the models relies to a degree on private institutions creating governance frameworks. Sometimes agencies will engage in rulemaking but adapt the substance from industry codes of conduct or

---

proper umbrella term. I have chosen the collaborative governance phrasing because it seems more descriptive.

174. *See, e.g.*, Emily S. Bremer, *Private Complements to Public Governance*, 81 MO. L. REV. 1115, 1123 (2016) (arguing that "private institutions [can] respond more nimbly, efficiently, and cost-effectively than administrative agencies to changes in technology, industry practice, or other circumstances").

175. Charles F. Sabel & William H. Simon, *Democratic Experimentalism*, *in* SEARCHING FOR CONTEMPORARY LEGAL THOUGHT 477, 489–93 (Justin Desautels-Stein & Christopher Tomlins eds., 2017).

176. *Cf.* Dennis D. Hirsch, *The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-Regulation*, 34 SEATTLE U. L. REV. 439, 466 (2011).

177. *See, e.g.*, Lobel, *supra* note 173, at 395–96 (praising new governance as a solution to law's fallibility).

178. *See* Jody Freeman, *Collaborative Governance in the Administrative State*, 45 UCLA L. REV. 1, 30 (1997).

179. *See id.*

180. *See* Cynthia Estlund, *Rebuilding the Law of the Workplace in an Era of Self-Regulation*, 105 COLUM. L. REV. 319, 341 n.94 (2005) ("The alternatives to command and control have many variations and varied names . . . . [A]ll of them involve some devolution of regulatory activity to the regulated entities themselves, all aim for greater flexibility, and all struggle with the tension between flexibility and accountability."); Lobel, *supra* note 173, at 345–47 (listing "reflexive law, soft law, collaborative governance, democratic experimentalism, responsive regulation, outsourcing regulation, reconstitutive law, post-regulatory law, revitalizing regulation, regulatory pluralism, decentering regulation, meta-regulation, contractarian law, communicative governance, negotiated governance, destabilization rights, cooperative implementation, interactive compliance, public laboratories, deepened democracy and empowered participatory governance, pragmatic lawyering, nonrival partnership, and a daring legal system" as theories that combine to form the concept of new governance). Scholars have not even agreed on the name of the field as a whole. *See* Estlund, *supra* (referring to the whole set as responsive regulation); CHRISTOPHER T. MARSDEN, INTERNET CO-REGULATION 59–63 (2011) (referring to the whole set as co-regulation); Kaminski, *supra* note 124, at 1559 (using "collaborative governance" and "new governance" interchangeably).

181. Douglas NeJaime, *When New Governance Fails*, 70 OHIO ST. L.J. 323, 332 (2009).

rely heavily on input from industry.[182] Other models are based on private governance structures with various check-ins, audits,[183] and monitoring[184] by the government. On the end closer to self-regulation, the government might primarily focus on its role as convener.[185] None of these models are mutually exclusive. In fact, one of the core insights of the collaborative governance movement is the importance of regulatory pluralism.[186] Scholars often speak in the metaphor of a regulatory toolkit, in which regulators can use the regulatory mode that is appropriate for the specific context.[187]

A regulation requiring companies to perform AIAs would be an example of a collaborative governance approach.[188] The regulation would require a private company to perform an AIA, which would then be overseen by a public regulator — a court or some form of

182. *See* Mendelson, *supra* note 173, at 739 ("The CFR today contains nearly 9,500 'incorporations by reference' of standards . . . . [M]any [rules] incorporate privately drafted standards from so-called 'standards development organizations' or 'SDOs,' organizations ranging from the American Society for Testing and Materials ('ASTM') to the Society for Automotive Engineers and the American Petroleum Institute ('API')."); IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION: TRANSCENDING THE DEREGULATION DEBATE 102 (1992); McGeveran, *supra* note 124, at 980; Freeman, *supra* note 178, at 33–55 (discussing examples of negotiated rulemaking); David A. Dana, *The New "Contractarian" Paradigm in Environmental Regulation*, U. ILL. L. REV. 35, 41 (2000) (discussing a contractarian model of regulation in which regulators and regulated entities negotiate on an enforceable set of requirements).

183. Douglas C. Michael, *Federal Agency Use of Audited Self-Regulation as a Regulatory Technique*, 47 ADMIN. L. REV. 171, 218–22 (1995); Kaminski, *supra* note 124, at 1535.

184. Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119 COLUM. L. REV. 369, 369 (2019).

185. Nick Doty & Deirdre K. Mulligan, *Internet Multistakeholder Processes and Techno-Policy Standards: Initial Reflections on Privacy at the World Wide Web Consortium*, 11 J. TELECOMM. & HIGH TECH. L. 135, 139 (2013) ("[T]he multistakeholder approach to privacy, which situates government as convener seeking to facilitate problem solving and the identification of consensus solutions among non-governmental experts, responds to perceived weaknesses of traditional 'command-and-control' regulation of the Internet consistent with general 'new governance' approaches to regulation."). Multi-stakeholder processes became particularly common tools for technology policy coordination in the Obama administration. *See id.*; Margot E. Kaminski, *When the Default Is No Penalty: Negotiating Privacy at the NTIA*, 93 DENV. L. REV. 925, 926 (2016) ("The federal government's current approach to data privacy concerns raised by these technologies is the underexamined multistakeholder process at the National Telecommunications and Information Administration (NTIA).").

186. *See, e.g.*, Neil Gunningham & Darren Sinclair, *Regulatory Pluralism: Designing Policy Mixes for Environmental Protection*, 21 L. & POL'Y 49, 50 (1999).

187. *See, e.g.*, Neil Gunningham, *Environmental Law, Regulation and Governance: Shifting Architectures*, 21 J. ENV'T L. 179, 179 (2009); Sarah E. Light, *The Law of the Corporation as Environmental Law*, 71 STAN. L. REV. 137, 212 (2019); Kaminski, *supra* note 124, at 1568 ("Collaborative governance emphasizes systemic accountability, or aggregate accountability, which looks at the interplay between different accountability mechanisms, over time.").

188. Kaminski & Malgieri, *supra* note 105, at 138 (describing the GDPR as an entire collaborative governance toolkit and arguing that the DPIA approach is one of the tools); Reuben Binns, *Data Protection Impact Assessments: A Meta-Regulatory Approach*, 7 INT'L DATA PRIV. L. 22, 29–30 (2017) (arguing that DPIAs are an example of meta-regulation, another category of new governance model).

regulatory monitor[189] — thus necessarily relying on a form of public-private partnership and enabling the use of technical and process knowledge that the government alone likely lacks.[190] The specific purposes of such a regulation line up with various strains of thought from collaborative governance discourse. The first AIA goal — to get system designers to change how they organize their production and planning processes — is an example of "reflexive regulation," regulation that "attempts to create incentives and procedures that induce entities to act in certain ways and to engage in internal reflection about what form that behavior should take."[191] The second goal — creating new information for later interventions — aligns both with theories of experimentalism[192] and policy-learning.[193] Experimentalism is the idea that better policy will come from many autonomous units trying to solve problems, coordinated by a central authority, based on the premise that this central authority could then update policies as it can determine what works.[194] Policy-learning is similar but more directly focused on designing policy to be alterable over time. The ability to update policy is central to that theory, while to experimentalists, those changes are a beneficial byproduct of otherwise important decentralized decisionmaking.[195] Here, each individual

189. Van Loo, *supra* note 184, at 397–98 (arguing that the rise of collaborative governance leads to the need for regulatory monitors); Jodi L. Short & Michael W. Toffel, *Making Self-Regulation More than Merely Symbolic: The Critical Role of the Legal Environment*, 55 ADMIN. SCI. Q. 361, 361 (2010) (arguing that heavy regulatory surveillance is important to fulfilling regulatory goals).

190. *See* Bremer, *supra* note 174, at 1123; Kaminski, *supra* note 124, at 1560.

191. Daniel J. Fiorino, *Rethinking Environmental Regulation: Perspectives on Law and Governance*, 23 HARV. ENV'T L. REV. 441, 447–48 (1999); *see also* David Hess, *Social Reporting: A Reflexive Law Approach to Corporate Social Responsiveness*, 25 J. CORP. L. 41, 51 (1999) ("Instead of directly regulating behavior to reach predetermined outcomes, reflexive law attempts to influence decision-making and communication processes with required procedures. The final decision, however, remains with the private actors."). Reflexive regulation is also similar to Cogliense & Lazer's "management-based regulation." Cary Coglianese & David Lazer, *Management-Based Regulation: Prescribing Private Management to Achieve Public Goals*, 37 L. & SOC'Y REV. 691, 692 n.1 (2003).

192. *See* Charles F. Sabel & William H. Simon, *Minimalism and Experimentalism in the Administrative State*, 100 GEO. L.J. 53, 79 (2011) (describing experimentalism as a regime where local units make autonomous decisions subject to coordination and monitoring by a central authority).

193. *See generally* Yair Listokin, *Learning Through Policy Variation*, 118 YALE L.J. 480 (2008) (describing approaches to policy that enable learning over time).

194. Sabel & Simon, *supra* note 192, at 78–80.

195. *See* Listokin, *supra* note 193, at 491 ("[Experimentalists] typically extol the virtues of federalism (and of other forms of decentralized decisionmaking) because of its learning benefits."); Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267, 316 (1998) (arguing that diverse sites will experience and solve similar problems and that this is a necessary precursor to large-scale improvements to current regulatory practice); Charles Sabel, Archon Fung & Bradley Karkkainen, *Beyond Backyard Environmentalism*, *in* BEYOND BACKYARD ENVIRONMENTALISM 3, 9, 13–14 (Joshua Cohen & Joel Rogers eds., 2000) (arguing that democratic experimentalism "dis-

technology firm preparing AIAs can demonstrate its different thought processes for mitigating algorithmic harms, and, collectively, we can learn from those differences.[196] Finally, the goal of bringing affected communities into the process is reflected in many forms of collaborative governance scholarship and practice. Early in the development of the literature, Ian Ayres and John Braithwaite wrote about what they called "tripartism" — the inclusion of public interest groups in new governance to help prevent capture,[197] and other scholarship has adopted different ideas for community involvement since.[198]

Understanding this regulation as a collaborative governance tool helps flesh out some of the choices that lawmakers would want to make about its contours. Collaborative governance prizes flexibility, suggesting that an AIA law would be better suited directing companies at the level of principles rather than specific actions.[199] The regulation should be forthright about which social goals it aims to implement, such as refraining from compounding injustice[200] or providing for effective recourse,[201] while offering the flexibility for the companies to explain how they designed their systems with those goals in mind, what choices they made, their varying approaches to

---

counts the possibility of central, panoramic knowledge" and does not make a "claim to a modest omniscience").

196. *See* Sabel & Simon, *supra* note 192, at 80 ("[E]xperimentalist regimes differ from command and control in that a large fraction of their norms are indicative or presumptive rather than mandatory. These regimes have mandatory norms requiring planning, reporting, monitoring, and, often, minimally satisfactory performance . . . . The function of nonmandatory rules is not to control discretion but to make practice transparent. They facilitate diagnosis and improvement.").

197. Ian Ayres & John Braithwaite, *Tripartism: Regulatory Capture and Empowerment*, 16 L. & SOC. INQUIRY 435, 435 (1991); AYRES & BRAITHWAITE, *supra* note 182, at 71.

198. *See generally* Lisa T. Alexander, *Stakeholder Participation in New Governance: Lessons from Chicago's Public Housing Reform Experiment*, 16 GEO. J. ON POVERTY L. & POL'Y 117, 128 (2009) (discussing and critiquing the ways that participation by marginalized groups is addressed in new governance scholarship); Freeman, *supra* note 178, at 30 ("[C]ommunity representatives may . . . join a company's internal quality circle"); *id.* at 32 (proposing "technical assistance grants or other needed support" for community group participation); *id.* at 82 (noting that agencies can "appoint[] a staff advocate or ombudsman for underrepresented groups"); Deborah N. Archer & Tamara C. Belinfanti, *We Built it and They Did Not Come: Using New Governance Theory in the Fight for Food Justice in Low-Income Communities of Color*, 15 SEATTLE J. FOR SOC. JUST. 307, 326–30 (2016) (proposing community-run tracking and monitoring scheme for food access issues).

199. Kenneth A. Bamberger & Deirdre K. Mulligan, *New Governance, Chief Privacy Officers, and the Corporate Management of Information Privacy in the United States: An Initial Inquiry*, 33 L. & POL'Y 477, 481 (2011) ("New governance approaches supplement, or sometimes replace, codified commands with more open-ended directives that leave significant discretion in their application").

200. *See* Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 105, 107–09 (Hugh Collins & Tarunabh Khaitan eds., 2018).

201. *See generally* Kaminski & Urban, *supra* note 85 (arguing for the importance of the right to contest AI decisions); Venkatasubramanian & Alfano, *supra* note 85, at 284–86 (arguing for the importance of recourse against algorithmic decisions).

mitigation, and what worked and did not. Oversight would not be a one-time affair, but an ongoing process with the goal of reaching a mutual understanding.

The complement to that flexibility, then, would be the need for an accountability framework to ensure meaningful cooperation and compliance. In general, accountability within collaborative governance systems operates by making compliance more attractive for the regulated entities; it is a regime of partnership rather than adversariality,[202] aiming to open "new lines of authority and accountability."[203]

As discussed in Part III, one common path for accountability is transparency, but the benefits of transparency here do not clearly outweigh the harms. Collaborative governance regimes do present the distinct possibility of capture,[204] which a degree of transparency is necessary to help mitigate.[205] But extreme transparency requirements are very likely to be actively resisted by the famously secretive technology companies,[206] both formally — with lobbying and public opposition to legislation — and informally, where internal firm actors may seek to sabotage the AIAs strategically to protect as much as they can without violating the law. This, presumably, is why DPIAs lack a transparency requirement[207] — its absence helps bring the firms to the table. From the perspective of AIAs as a collaborative governance mechanism, it is not obvious that transparency is necessary, or entirely a net positive. But if transparency is not the answer, then something would certainly need to replace it from an accountability perspective.

There are several ideas about accountability in the new governance literature that revolve around structuring incentives in clever ways. The traditional approach to ensuring an AIA is performed would be to penalize failure to do so. But that does not offer any sort of guarantee of the quality of the AIA. It also works against the collaborative spirit of the regulation and would lead companies to treat the AIA merely as a form to fill out. Instead, collaborative governance approaches tend to favor ranges of encouragement and sanctions that may or may not be formal. The ideal form of this approach would get companies so invested in the success of the regulation itself through that collaborative process that they will want to cooperate — what

202. *See, e.g.*, David Hess, *Social Reporting: A Reflexive Law Approach to Corporate Social Responsiveness*, 25 J. CORP. L. 41, 64 (1999) (internal quotation marks and footnote omitted) ("The goal of reflexive regulation is not to cause corporations to engage in defensive compliance, but to encourage proactive, socially responsive management.").

203. Freeman, *supra* note 178, at 30.

204. David Thaw, *Enlightened Regulatory Capture*, 89 WASH. L. REV. 329, 336 (2014) ("Engaging private expertise does, however, carry substantial risk of regulatory capture.").

205. *See* Margot E. Kaminski, *Understanding Transparency in Algorithmic Accountability*, *in* THE CAMBRIDGE HANDBOOK OF THE LAW OF ALGORITHMS, *supra* note 1, at 121, 127; Hirsch, *supra* note 176, at 468.

206. *See* Katyal, *supra* note 48, at 1193.

207. *See* DPIA Guidance, *supra* note 116, at 18.

David Thaw terms "enlightened regulatory capture."[208] But many other possibilities exist for structuring incentives. A typical form of encouragement is something like a safe harbor, though safe harbors also run the risk of giving too much immunity too easily.[209] On the penalty side, scholars suggest not a single penalty, but escalating penalties, such that companies would cooperate in their own punishment for a violation to avoid the lurking maximum penalty.[210]

Another idea that occurs with some frequency is that of the "penalty default."[211] The idea of a penalty default is a regulatory provision that makes the default outcome bad enough that companies will want to cooperate to get around it.[212] The difference between an imposed penalty and a penalty default is that the default — as the phrase suggests — kicks in automatically, without any need for the regulator to choose to apply it. This helps to situate the regulator as a partner rather than an adversary, preserving the cooperative spirit of the arrangement and allowing collaboration to continue.

Penalty defaults can be fines, but other options exist. The specter of regulation can itself act as a penalty default. This is a familiar story in the arena of self-regulation; as soon as regulation seems likely, companies find they should come together and hash out a code of conduct to hold themselves to. Dennis Hirsch argued that the United States should pass a baseline privacy law for just this reason.[213] Alternatively, Kristelia García has argued that because certainty is highly valued in the private sector, the continued cost of legal uncertainty can act as a penalty default.[214] Finally, Bradley Karkkainen argued in the NEPA context that the expense of an impact statement itself can be a penalty default.[215] Because NEPA allows firms to skip a full environmental impact statement ("EIS") if they match their environmental practices to previously accepted solutions, the cost of a full EIS can act as a penalty default to encourage substantive compliance early

---

208. *See* Thaw, *supra* note 204, at 332 (capitalization omitted). This approach ties in with new governance's beneficial take on managerialism, as discussed in Section III.B.

209. *See generally*, Pauline T. Kim, *Safe Harbors for Algorithms?* (unpublished manuscript) (draft on file with author).

210. AYRES & BRAITHWAITE, *supra* note 182, at 43–44. The GDPR's maximum penalty of 4% of a regulated entity's global revenue is a great example of a "super-punishment" that allows regulators to more easily impose lesser fines. *See id.*; Commission Regulation 2016/679 of Apr. 27, 2016, art. 83(5), 2016 O.J. (L 119) 83.

211. Kaminski, *supra* note 185, at 926; Karkkainen, *supra* note 33, at 903; Sabel & Simon, *supra* note 175, at 496–97.

212. *See* Kristelia A. García, *Penalty Default Licenses: A Case for Uncertainty*, 89 N.Y.U. L. REV. 1117, 1122 (2014) (discussing the origins of the term in contract theory).

213. Dennis D. Hirsch, *Going Dutch? Collaborative Dutch Privacy Regulation and the Lessons it Holds for U.S. Privacy Law*, 2013 MICH. ST. L. REV. 83, 159 (2013) (arguing that "a baseline privacy statute" would "provide a structure for the industry codes and to give companies a strong incentive to come to the table and negotiate a code of conduct").

214. García, *supra* note 212, at 1121.

215. Karkkainen, *supra* note 33, at 936.

on.[216] An AIA statute could leverage such an idea with similarly graded compliance requirements: if a firm develops its technology in ways that have been judged adequate in prior projects, perhaps it can avoid a costly repeat of the full AIA.[217]

While there are known ways to encourage better compliance with filling out the AIA, meaningfully involving affected communities will be more difficult. Though many of the specific AIA proposals envision input from affected communities,[218] none of the proposals has much detail or a workable vision for how to accomplish it,[219] nor a way to go beyond mere input to some form of power over the system.[220] The NEPA model relies on formal notice and comment,[221] but a single chance to comment does not allow for the back-and-forth problem-solving stance that underlies the whole concept of collaborative governance. In addition, as Jody Freeman has pointed out, community and public interest groups often lack both the technical and legal expertise necessary to engage with a formal process.[222] Finally, notice and comment would require a more absolutist stance on transparency that might hinder cooperation, as discussed above. The DPIA model is not obviously better, though, as it allows companies free reign to get community input if and when they deem it necessary. Private firms do have technical expertise, but they do not have expertise in determining what communities will be affected adversely by their product nor do they have expertise in soliciting and synthesizing input from those communities. The DPIA model could work in spirit but it would need more oversight than it currently receives.

Collaborative governance scholarship does touch on methods for involving community input, but it is unsatisfying. The need for community input is regularly noted followed by a list of challenges, such as the need for assistance and funding to meaningfully participate, the

---

216. *Id.*

217. Selbst*, supra* note 37, at 185.

218. *Id.* at 178; Kaminski & Malgieri, *supra* note 105, at 139.

219. Outside of the AIA proposals specifically, machine learning researchers have begun to think about participatory design processes for community involvement. *See* P. M. Krafft et al., *An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 772, 772 (2021) (proposing the Algorithmic Equity Toolkit). But participatory design is not a panacea either. *See* Christina N. Harrington, Sheena Erete & Anne Marie Piper, *Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements*, 3 PROC. ACM CONF. ON HUM.-COMPUT. INTERACTION, 216, 216 (2019) (critiquing participatory design as "an affluent and privileged activity that often neglects the challenges associated with envisioning equitable design solutions among underserved populations").

220. *See* K. Sabeel Rahman & Jocelyn Simonson, *The Institutional Design of Community Control*, 108 CALIF. L. REV. 679, 719–22 (2020); Mona Sloane, Emanuel Moss, Olaitan Awomolo & Laura Forlano, *Participation Is Not a Design Fix for Machine Learning*, 119 PROC. MACH. LEARNING RSCH. (2020).

221. *See supra* note 115 and accompanying text.

222. *See* Freeman, *supra* note 178, at 80.

lack of methods for deciding what the relevant groups are and who gets to represent them, private firms' resistance to including other stakeholders, and adding length and complexity to the project planning process.[223] These challenges are typically presented without accompanying solutions. The SIA literature has described similar challenges as well. As Ana Maria Esteves, Daniel Franks, and Frank Vanclay observed in the latest "state of the art" article:

> The adequacy of public participation continues to be an issue. SIAs often do not meet public expectations of being a deliberative process to determine the acceptability of a project. Rather they are seen at best as a process for incremental project improvement, and at worst as being little more than a feeble attempt at project legitimization. Public participation ranges from being the provision of periods for public comment and the supply of information, to being the active involvement of stakeholders in shaping the SIA process and the opening-up of governance processes to include local communities in decisionmaking about projects.
>
> The demands of community consultation can lead to fatigue in communities and local governments, particularly in situations with multiple developments. These challenges are exacerbated where there is limited engagement, leading participants to question the value of their involvement.[224]

The concerns with algorithmic harms will constantly fall on people who currently lack input into the process, let alone any power over the systems, even as algorithmic technologies drastically alter their lives. If this regulation is to make any difference, there must be an avenue for communities to have input and eventually a measure of control.[225]

---

223. *See, e.g.*, *id.*; Dana, *supra* note 182, at 54–55; Alexander, *supra* note 198, at 137–38.

224. Esteves et al., *supra* note 35, at 37.

225. Some work has been done on involving communities in the governance of technology that suggests paths forward. *See* Citron, *supra* note 9, at 1312 (proposing that "information technology review boards that would provide opportunities for stakeholders and the public at large to comment on a system's design and testing"); Min Kyung Lee et al., *WeBuildAI: Participatory Framework for Algorithmic Governance*, 3 PROC. ACM CONF. ON HUM.-COMPUT. INTERACTION 181, 181 (2019) (proposing a participatory framework for algorithmic governance).

B. Legal Managerialism

Collaborative governance methods are primarily, if not totally, procedural in nature.[226] They treat the challenges of regulation as those of ensuring better process through cooperation, information, convenings, and diversity, but tend to remain agnostic as to substantive outcomes. This procedural nature is a characteristic of impact assessments. Though NEPA was passed with forcefully stated substantive goals for environmental improvement, those goals were unspecific about outcomes.[227] When the law faced a hostile Supreme Court,[228] those substantive provisions were eventually held to be irrelevant.[229] Though it would be possible in theory to write AIA requirements to be more specific on substance, the Court's gloss on NEPA helped solidify the idea within American legal culture that impact assessments as a method are primarily procedural.[230] We therefore do not have great answers about how to combine those procedures with substantive regulatory goals. Moreover, we cannot rely on statutory language to do so without running the risk that a

---

226. *See* Eric W. Orts, *Reflexive Environmental Law*, 89 NW. U. L. REV. 1227, 1264 (1995) ("Reflexive solutions offload some of the weight of social regulation from the legal system to other social actors. This is accomplished by *proceduralization*.").

227. The statute declares that the government is to use "all practicable means, consistent with other essential considerations of national policy," to "maintain conditions under which man and nature can exist in productive harmony, and fulfill the social, economic, and other requirements of present and future generations of Americans." National Environmental Policy Act of 1969 § 102(2)(A), 42 U.S.C. § 4331(a).

228. Several commentators have observed that NEPA never had a chance to operate as written because of the Supreme Court's interpretations. *See, e.g.*, Philip Michael Ferester, *Revitalizing the National Environmental Policy Act: Substantive Law Adaptations from NEPA's Progeny*, 16 HARV. ENV'T L. REV. 207, 217–23 (1992) (discussing the Supreme Court's jurisprudence on NEPA in a section entitled "NEPA Before the Supreme Court: Extinguishing Substantive Review"); David R. Hodas, *NEPA, Ecosystem Management and Environmental Accounting*, 14 NAT. RES. & ENV'T 185, 186 (2000) (arguing that the Supreme Court has been too deferential to agencies, thus undermining the substantive goals of NEPA); Nicholas C. Yost, *NEPA's Promise — Partially Fulfilled*, 20 ENV'T L. 533, 534 (1990) (arguing that "[s]ubstantive review under NEPA" is "essentially unfulfilled"); *The National Environmental Policy Act: An Interview with William Hedeman, Jr.*, EPA (1980), https://archive.epa.gov/epa/aboutepa/national-environmental-policy-act-interview-william-hedeman-jr.html [https://perma.cc/Y2KR-TYZJ] ("I feel that much of NEPA's problem in the past has been the manner in which it has been interpreted by the courts . . . . Unfortunately, most of this litigation has focused on procedural compliance with the requirements of NEPA rather than getting to the basic substantive mandates of the Congress as reflected in NEPA's goals and policies.").

229. *See* Strycker's Bay Neighborhood Council, Inc. v. Karlen, 444 U.S. 223, 227 (1980) (overturning the Second Circuit's holding that an agency's environmental determinations "should be given determinative weight," ruling that NEPA is "essentially procedural," and not subject to even arbitrary and capricious substantive review); *see also* Yost, *supra* note 228, at 540 ("[T]he Supreme Court . . . early employed unduly restrictive dicta to characterize NEPA's role and then became the captive of its own earlier dicta.").

230. *See* sources cited *supra* note 228. There are still some exceptions, however. *See* Katyal, *supra* note 48, at 115 (noting that California's state environmental impact legislation contains substantive requirements).

court would simply ignore the substantive goals again. Thus, we should assume that AIA regulations will be primarily procedural.

When procedural regulations are implemented by a regulated firm, the firm's priorities will shape the implementation. The culture and structures of organizations are totalizing forces that reshape the goals of every action or plan that comes under the organization's control.[231] In the private sector, this means that policies will be reframed in terms of how they affect efficiency and profit.[232] These tendencies have been variously called the "managerialization of law"[233] or managerialism more generally, applied in the context of law.[234] Here, I refer to the concept broadly as legal managerialism.

Managerialist tendencies have been documented across different legal contexts. Lauren Edelman has shown in the context of employment that firms tend to speak more about diversity than discrimination, because they can more easily offer business-related reasons to support diversity than to prevent discrimination.[235] Scholars have observed similar dynamics in privacy law. Kenneth Bamberger and Deirdre Mulligan found that companies measured privacy by reference to the need to satisfy consumer expectations.[236] Ari Waldman found that companies treat data security more seriously than consumer privacy because the impact on the bottom line is clearer.[237] In the algorithmic context specifically, Dennis Hirsch and colleagues have

---

231. *See* COHEN, *supra* note 143, at 244; Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State,* 56 DUKE L.J. 377, 383 (2006) ("The[] problems [of delegating regulation to private firms] are rooted not in self-interested calculation about private gain or shortcomings in normative commitments to legal compliance, but in the less conscious workings of organizational decision processes.").

232. Bamberger, *supra* note 231, at 383 ("[T]hese pathologies are especially pronounced when regulatory norms cause a drag on efficiency, i.e., when those norms are in tension with the core goals around which the firm is structured."). The idea applies equally to the public sector, where policies are reframed in terms of the agency's primary mission. *See* Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance by Design*, 106 CALIF. L. REV. 697, 701–02 ("Constrained by mission and statute, individual agencies possess neither the constitutional ability nor the structural incentives to consider competing values outside their narrow ambit. Such agency-by-agency decisionmaking creates downstream ripple effects, prioritizing certain values and precluding reasoned deliberation over others."); J.R. DeShazo & Jody Freeman, *Public Agencies as Lobbyists*, 105 COLUM. L. REV. 2217, 2219–20 (2005).

233. Lauren B. Edelman, Sally Riggs Fuller & Iona Mara-Drita, *Diversity Rhetoric and the Managerialization of Law*, 106 AM. J. SOCIO. 1589, 1592 (2001) (defining the "managerialization of law" as "the process by which conceptions of law may become progressively infused with managerial values as legal ideas move into managerial and organizational arenas").

234. COHEN, *supra* note 143, at 144.

235. LAUREN B. EDELMAN, WORKING LAW: COURTS, CORPORATIONS, AND SYMBOLIC CIVIL RIGHTS 142–46 (2016).

236. KENNETH A. BAMBERGER & DEIRDRE K. MULLIGAN, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE 65–66 (2015).

237. *See* Ari Ezra Waldman, *Designing without Privacy*, 55 HOUS. L. REV. 659, 712 (2018).

found that companies are motivated by reputation and consumer trust as much as by social goals.[238]

So what are the consequences of managerialism for a law's effectiveness? Where a classic adversarial view of the regulator in a struggle to rein in the regulated will see this as conflict, scholars who are oriented toward collaborative governance approaches may take the fact of managerialism as a starting point and organize their thinking around it.[239] If policy can succeed at aligning substantive regulatory goals with profit and efficiency, then the profit motive and organizational logics can be used for good as companies can embed the social concerns into their profit motive.[240] For example, Bamberger and Mulligan argue that embedding privacy concerns within corporate risk analyses would cause companies to take privacy more seriously.[241] Specifically, managerialism creates distance between the people in the company who set privacy policy and the everyday practice of it, which functions better when it is seen as "an apolitical business requirement."[242] Collaborative governance approaches actually rely on managerialism: they seek to encourage compliance and "beyond compliance" behaviors,[243] and it only makes sense that aligning social goals with company incentives would help accomplish that end.

Legal scholarship skeptical of private governance paints a darker picture of managerialism. Lauren Edelman's large body of work about how firms construct responses to anti-discrimination law is typical of this skeptical picture. Edelman has argued that the ambiguity of legal requirements leaves firms opportunities to interpret and co-construct them, resulting in what she refers to as legal endogeneity.[244] Edelman describes a six-stage cycle in which ambiguous legal requirements allow firms to endogenously construct the meaning of laws and create compliance structures to fulfill their legal obligations, such as human resources offices, diversity officers, and trainings.[245] These compliance structures are often hollow, frustrating the substantive purpose of

238. DENNIS HIRSCH, TIMOTHY BARTLEY, ARAVIND CHANDRASEKARAN, DAVON NORRIS SRINIVASAN PARTHASARATHY & PIERS NORRIS TURNER, BUSINESS DATA ETHICS: EMERGING TRENDS IN THE GOVERNANCE OF ADVANCED ANALYTICS AND AI 27–32 (2021) (detailing empirical research on AI ethics showing that many companies are trying to manage AI harms to improve reputation and anticipated competitive advantage).

239. *See, e.g.*, Coglianese & Lazer, *supra* note 191, at 702–03, 711 (arguing that the management-based regulation can be a beneficial approach where managerial incentives align with social goals, and the degree to which regulation should require mandates should be based on how much the social goal aligns with preexisting managerial incentives).

240. BAMBERGER & MULLIGAN, *supra* note 236, at 177 (discussing how managerialization of privacy law means that firms take privacy law more seriously than they otherwise would have).

241. *Id.* at 177–78.

242. *Id.* at 178.

243. *See infra* Section IV.C.

244. EDELMAN, *supra* note 235, at 27–41.

245. *Id.*

the legal requirements, but nonetheless end up serving as stand-ins for substantive compliance with the law.[246] Finally, the process ends with judges looking to the symbolic compliance structures as evidence of compliance with the law, even if the laws' purposes — such as workplace equality — remain entirely unfulfilled.[247] Waldman, drawing on Edelman's work, finds similar trends in privacy law: procedural legal requirements meant to vindicate substantive rights, which are then hollowed out in implementation.[248]

In Edelman's account, managerialism at least partly explains why compliance structures fail to vindicate their substantive policy goals. For example, instead of speaking about discrimination and justice, compliance professionals speak in terms of risk and litigation exposure, or how diversity can improve profits.[249] This account of managerialized law differs from the collaborative governance scholarship: it argues that firms do not simply incorporate normative and legal mandates into business practices, but rather that the process of doing so warps them into frameworks that fit the existing organizational logics of the firm. Though not every instance of legal interpretation done inside a firm will be ambiguous, there will be many occasions in which legal and business ideals either are or appear to be in conflict. The managerialization process then ensures that business ideals win out over justice or fairness ideals in those conflicts.[250] The end result is that in many cases, the substantive legal goals that firms are meant to comply with go unrealized because the compliance structures hollow out the substantive legal requirement in favor of preexisting organizational goals.

There is good reason to believe that private sector impact assessment requirements will also turn out to be legally endogenous. AIA legislation must be somewhat ambiguous in its requirements because the approach is defined partly by its flexibility and its reliance on firm expertise. Given that priorities for private sector tech firms are the same as any other firm, all of Edelman's other factors are present as well. Indeed, where impact assessments are voluntary, their proponents often demonstrate a managerial frame by focusing on costs and business risk.[251] Unsurprisingly, practitioners of social impact assessments note that impact assessments are sometimes tailored to

246. *Id.*
247. *Id.*
248. Waldman, *supra* note 148, at 806–10.
249. EDELMAN, *supra* note 235, at 142–49.
250. *Id.* at 33.
251. David Wright, *The State of the Art in Privacy Impact Assessment*, 28 COMPUT. L. & SEC. REV. 54, 55 (2012) ("[A] PIA helps reduce costs in management time, legal expenses and potential media or public concern by considering privacy issues early. It helps an organisation to avoid costly or embarrassing privacy mistakes.").

meet only minimum expectations of regulators.[252] And just as judges were willing to accept symbolic structures as proof of compliance in the Title VII context, so too have they shown themselves willing to treat impact assessments as purely procedural requirements in the past.[253]

Rather than an explanation of the inner workings of organizations, to scholars like Julie Cohen, managerialism is a manifestation of neoliberal ideology external to the firm that subjugates social policy to market norms and logics more generally.[254] Cohen points to the managerialist trends in dispute resolution that prize efficiency over justice, and as a result, warp the very judicial systems that one might otherwise have assumed would be fixed.[255] Whereas Edelman's account might suggest that collaborative governance techniques are doomed to failure, for Cohen the neoliberal managerialist frame is intellectually prior to — and an explanation for — the popularity and utility of collaborative governance approaches to regulation. Rather than doomed to fail, collaborative governance approaches are a product of a failed legal imagination that begins by accepting the primacy of market objectives in the first place.[256] In practical consequence, these two critiques are not far from each other. Each suggests that firms take in substantive legal requirements and churn out reputation-enhancing compliance structures that may ultimately offer little of substance as to the original motivating concern.[257]

Though not situated within a legal compliance regime, concerns with managerialism are already being validated in this industry by the rise of "AI ethics." Researchers who have examined AI ethics have found that within technology companies, "ethics" has been subsumed by the business logics inherent in the technology companies that seek to self-impose ethical codes.[258] Indeed they appear almost to mirror

252. Esteves et al., *supra* note 35, at 36 ("The limited capacity of regulators and the limited resources devoted to quality control have a significant impact on the standard of SIAs, with a tendency for proponents to produce assessments that only just pass the minimum expectations of regulators.").

253. *See supra* note 228 and accompanying text.

254. COHEN, *supra* note 143, at 143–47, 154–55 (describing managerialism as a "form of institutional discipline that has gradually but inexorably swept the judicial system into its orbit").

255. *Id.*

256. *See* Corinne Blalock, *Neoliberalism and the Crisis of Legal Theory*, 77 L. & CONTEMP. PROBS. 71, 90 (2014) (arguing that neoliberalism's hegemony and associated invisibility in legal discourse is an important source of its influence).

257. COHEN, *supra* note 143, at 245–46 (discussing the effects of managerialism on human rights efforts).

258. Jacob Metcalf, Emanuel Moss & danah boyd, *Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics*, 86 SOC. RSCH.: AN INT'L Q. 449, 455 (2019) ("[A]s ethical product design and governance goals are becoming institutionalized by tech firms, the practices associated with these goals are being crafted and executed according to the existing logics and structures of the technology industry, even as they are responding to outside critiques of these logics and structures."); *see also* Elettra Bietti, *From*

the steps Edelman describes for legal endogeneity. First comes risk framing, where "ethics owners" within companies try to grapple with the difficult social questions, but face steep financial pressure from investors to think of ethics only as a necessity to avoid "downside risk," and to "implement [only] ethics practices that do not negatively affect companies' bottom lines."[259] Next comes compliance structures such as creation of checklists and toolkits,[260] hiring of ethics officers,[261] or creating ethics oversight boards.[262] These compliance structures can be substantive or symbolic, but nothing about their existence ensures either. Then comes managerialization, where moral questions about how the technologies and technology companies interact with the world around them become restructured as problems with the technology itself that can be answered with better technical design.[263] These are solutions that fit with the logics of a technology company, which in the best cases double as products that the company itself can sell. Finally, ethics structures are mobilized into a movement to gain further legitimacy via "ethics washing" in response to public backlash against dominance by technology firms.[264] There is some evidence that a similar cycle is starting with the nascent algorithmic auditing industry as well.[265]

One need not be cynical about these co-optations of ethics to appreciate how they happen. Rather, it is enough to recognize that to the technology companies, ethics is outside of their core competence and is at best a secondary mission.[266] They may hire ethicists, but funda-

---

Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 210, 210 (2020) (arguing that ethics as a method should not simply be equated with technology companies' desires for self-regulation).

259. Metcalf et al., *supra* note 258, at 465.

260. *See supra* Section II.C.

261. *Rise of the Chief Ethics Officer*, FORBES (Mar. 27, 2019, 1:22 PM), https://www.forbes.com/sites/insights-intelai/2019/03/27/rise-of-the-chief-ethics-officer/#797784835aba [https://perma.cc/B8CL-63ZX].

262. James Vincent, *The Problem with AI Ethics*, VERGE (Apr. 3, 2019, 11:47 AM), https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech [https://perma.cc/RJ5L-TBEX].

263. Daniel Greene, Anna Lauren Hoffmann & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, 52 PROC. HAW. INT'L CONF. ON SYS. SCI. 2122, 2129 (2019); *see also* Selbst et al., *supra* note 3, at 63 (discussing the "Solutionism Trap").

264. *See, e.g.*, Hao, *supra* note 146.

265. *See* Ng, *supra* note 38.

266. *See* Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan & Hanna Wallach, *Co-designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI*, PROC. CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1, 6 (2020) ("The disconnect arising from rhetorical support for AI fairness efforts coupled with a lack of organizational incentives that support such efforts is a central challenge for practitioners."); *id.* at 9 ("[W]ithout AI fairness efforts 'moving any of the top-line metrics,' [practitioners] feel unable to properly justify the resources needed to address issues, given their other priorities during the development and deployment lifecycle.").

mentally, the organizations' structures, attitudes, company cultures, and experiences are set up to build technology for profit, while pressures of the market are real and enforced by shareholders. If these processes can operate to undermine existing legal requirements in the absence of any intent to do so, it only makes sense that they can similarly erode self-regulatory public commitments.

What lessons can legislators take from these critiques? The strong form of Cohen's account might suggest that we just give up, that impact assessments that give discretion to companies are simply the wrong approach to regulation. But I don't take either Cohen's or Edelman's arguments to necessitate that conclusion. I believe the best response for AIA legislation would be to resist neoliberal framing to the extent possible by demanding certain minimum substantive outcomes, even — and perhaps especially — where they are irreconcilable with a profit motive.

This may prove a conceptual challenge. There is a tension between substantive and procedural standards when it comes to enforcement. At some point, if a reviewing body has the authority to look at an impact assessment that was procedurally proper and reject it for failing to address the substantive harms well enough, then the review is, in practice, substantive, not procedural. This tension is partially to blame for the fate that befell NEPA's original substantive provisions, where the Supreme Court framed any substantive review by a lower court as "substitut[ing] its judgment for that of the agency."[267] But such a claim of substitution of the court's judgment is only coherent where the court has to *use* judgment — that is, when the value statements in the statute are vague or ambiguous. If the substantive demands are instead spelled out clearly and concretely in statute or regulation, then this tension can be resolved. This would resolve legal endogeneity concerns as well because without ambiguity, the legal endogeneity cycle never begins.[268] The danger of legal managerialism is thus tied directly to ambiguity — the law can only be undermined where there is room to maneuver and interpret. Given the inherent desire for flexibility in a collaborative governance scheme, this suggests that AIA legislation should dictate minimum substantive standards for desired outcomes, while treating everything in excess as a governance problem.[269]

---

267. *See* Kleppe v. Sierra Club, 427 U.S. 390, 410 n.21 (1976) (describing substantive review by courts as "substitut[ing] its judgment for that of the agency"), *quoted in* Vermont Yankee Nuclear Power Corp. v. Nat. Res. Def. Council, Inc., 435 U.S. 519, 555 (1978).

268. Waldman*, supra* note 41, at 17.

269. *See* Freeman, *supra* note 178, at 32 ("Agency officials need not be agnostic about outcomes. They may limit the universe of subjects open to negotiation by establishing, for example, a . . . minimal floor, while leaving to a consensus-based process how high above that minimum the standard should ultimately be set in light of feasibility, cost effectiveness, and community priorities.").

Minimum substantive standards present an additional advantage and disadvantage. The advantage is that they are easy to oversee and administer. The more concrete the standards are, the easier and less costly they are to check. This is even more important when one considers that "impacts" that will be assessed are not the same as harms; rather, in the words of Jacob Metcalf and colleagues, impacts "are constructs that act as proxies for the often conceptually distinct sociomaterial harms algorithmic systems may produce."[270] Metcalf and colleagues point out that AIAs embed a threshold question of what harms even get counted as impact worth discussing.[271] This threshold must not simply be left to the discretion of the assessor, and a minimum substantive standard in the AIA legislation or implementing regulations should specify which types of harms count.

The disadvantage is that the more that the statute relies on compliance with concrete standards — and the necessary oversight that entails — the more companies will treat compliance as the goal and the regulator as adversarial, rather than seeing the goal as cooperatively solving algorithmic harms. This can be a serious downside if it negates one important benefit of a cooperative approach: encouraging "beyond compliance" behaviors, the subject of the next Section.

C. Beyond Compliance Behaviors

The compliance behavior of firms in the face of new or existing regulation varies widely.[272] The managerialist phenomenon should temper expectations that individual firms will *all* substantively comply, but there is a different phenomenon that can serve as a counterweight: "beyond compliance" behaviors. Beyond compliance behaviors occur where firms devote more resources and efforts to public goals than is strictly required by law. Examples of beyond compliance behaviors include pulp mills spending significant resources to combat foul odors, despite not being legally required to,[273]

---

270. Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh & Madeleine Clare Elish, *Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 735, 736 (2021).

271. *See id.* at 737–40.

272. *See, e.g.*, Denise L. Anthony, Ajit Appari & M. Eric Johnson, *Institutionalizing HIPAA Compliance: Organizations and Competing Logics in U.S. Health Care*, 55 J. HEALTH & SOC. BEHAV. 108, 109–10 (2014) (summarizing the literature); Thomas D'Aunno, Melissa Succi & Jeffrey A. Alexander, *The Role of Institutional and Market Forces in Divergent Organizational Change*, 45 ADMIN. SCI. Q. 679, 682 (2000); Robert A. Kagan, Neil Gunningham & Dorothy Thornton, *Explaining Corporate Environmental Performance: How Does Regulation Matter?*, 37 L. & SOC'Y REV. 51, 51–52 (2003).

273. Neil Gunningham, Robert A. Kagan & Dorothy Thornton, *Social License and Environmental Protection: Why Businesses Go Beyond Compliance*, 29 L. & SOC. INQUIRY 307, 322, 328 (2004).

or Apple updating its mobile operating system to require that apps request permission from customers to track them, even without any regulation requiring it.[274]

Scholars have tried to explain why some firms appear to go beyond pure legal compliance. Some behaviors can still be attributed to long-term profit incentives. Companies may figure that even if something is not regulated yet, it will be in the future, and they can gain a competitive advantage by being ahead of the curve.[275] Beyond compliance behaviors lead to benefits in corporate reputation, which can also be linked to future profits if consumers shop based on social considerations over market ones.[276] This is a managerialist explanation for the phenomenon; managerialism doesn't prevent companies from doing good things if they perceive profit in it.

Other motivations exist as well. Neil Gunningham, Robert Kagan, and Dorothy Thornton have argued that firms must act within their "social license" to operate, and "the interplay between social pressures and economic constraints" better explains beyond compliance behavior.[277] In a study of the trucking industry, they found that in the absence of regulation, "social and normative pressures for better environmental performance are likely to be minimal in highly competitive industries," but that companies with "large truck fleets and widely recognizable names and consumer reputations at stake" will more likely take beyond compliance measures.[278] This suggests that when company survival appears to be at stake, firms are coldly rational about competition and efficiency, but when that pressure lifts there may be room to focus more on beneficence.

Another factor is who is employed and empowered by the firm. In his study of the "greening" of different firms, Aseem Prakash found that leaders within firms matter, along with the extent to which they

---

274. *See* Kate O'Flaherty, *Apple's Stunning IOS 14 Privacy Move: A Game-Changer for All iPhone Users*, FORBES (Jan. 31, 2021, 3:59 AM), https://www.forbes.com/sites/kateoflahertyuk/2021/01/31/apples-stunning-ios-14-privacy-move-a-game-changer-for-all-iphone-users [https://perma.cc/NS2M-UJXB].

275. Gunningham et al., *supra* note 273, at 308.

276. The evidence on whether consumers do change their behavior based on corporate ethics is somewhat mixed. *See, e.g.*, Lois A. Mohr, Deborah J. Webb & Katherine E. Harris, *Do Consumers Expect Companies to be Socially Responsible? The Impact of Corporate Social Responsibility on Buying Behavior*, 35 J. CONSUMER AFFS. 45, 49–50 (2001) (finding in survey research that a small but committed percentage of consumers based purchasing decisions on social practices in part); Alexander Chernev & Sean Blair, *Doing Well by Doing Good: The Benevolent Halo of Corporate Social Responsibility*, 41 J. CONSUMER RSCH. 1412, 1412 (2015) (finding that social goodwill towards a company can affect consumers' perceptions of their products). Nonetheless, what matters more is whether firms believe that beyond compliance behaviors influence consumer behavior and thus profit.

277. Gunningham et al., *supra* note 273, at 307.

278. Dorothy Thornton, Robert A. Kagan & Neil Gunningham, *When Social Norms and Pressures Are Not Enough: Environmental Performance in the Trucking Industry*, 43 L. & SOC'Y REV. 405, 407, 408–10 (2009).

view environmental concerns as important on their own merits.[279] Similarly, one caveat that Edelman offered in her explanation of legal endogeneity recognizes the importance of individual activists within firms:

> In any given organization, symbolic structures may be more or less effective at promoting legal ideals. Across organizational fields . . . managerialization collectively weakens the capacity of law to overcome business practices that frustrate legal ideals . . . . By contrast, when compliance professionals are committed to legal ideals and adopt an activist stance, they can render symbolic structures more substantive.[280]

Such "institutional entrepreneurs" help drive the social agenda of a company.[281] Early empirical research into AI ethics within companies echoes the importance of internal activism. As Michael Madaio and colleagues found in their research with AI ethics industry practitioners, "AI fairness efforts are often the result of ad-hoc processes, driven by passionate individual advocates."[282] Thus the identification of activists in-house and management willing to grant them room to run may turn out to be signals that the firm will end up being a leader on preventing algorithmic harms.

The neoinstitutionalist school within organizational sociology stresses the importance of pressures on institutions within a field.[283] Central to this line of thought is the phenomenon of "institutional isomorphism," popularized in a formative article by Paul DiMaggio and Walter Powell.[284] DiMaggio and Powell argue that rather than firms being rational profit-maximizers driven to change by the need

---

279. ASEEM PRAKASH, GREENING THE FIRM: THE POLITICS OF CORPORATE ENVIRONMENTALISM 145–47 (2000); *see also* Madaio et al., *supra* note 266, at 9 (describing the influence of a CTO deciding to make AI ethics a priority).

280. EDELMAN, *supra* note 235, at 36.

281. *See generally* Paul J. DiMaggio, *Interest and Agency in Institutional Theory*, *in* INSTITUTIONAL PATTERNS AND ORGANIZATIONS 3, 3–22 (Lynne G. Zucker ed., 1988). When they have influence, institutional entrepreneurs likely direct institutional priorities in a way that is beneficial to their own interests, *see* Jens Beckert, *Agency, Entrepreneurs and Institutional Change: The Role of Strategic Choice and Institutionalized Practices in Organizations*, 20 ORG. STUD. 777, 781 (1999), but the important point is that sometimes such interests will align with policy goals.

282. Madaio et al., *supra* note 266, at 2.

283. *See generally* THE NEW INSTITUTIONALISM IN ORGANIZATIONAL ANALYSIS (Walter W. Powell & Paul J. DiMaggio eds., 1991).

284. Paul J. DiMaggio & Walter W. Powell, *The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields*, 48 AM. SOCIO. REV. 147, 147 (1983); *see also* John W. Meyer & Brian Rowan, *Institutionalized Organizations: Formal Structure as Myth and Ceremony*, 83 AM. J. SOCIO. 340, 340 (1977).

for ever greater efficiency, firms in reality start out with important differences in organizational attitudes and individual leadership, and then homogenize over time due to social processes that have nothing to do with efficiency or profit.[285] Institutional isomorphism operates through three mechanisms: coercive, mimetic, and normative.[286] Coercive isomorphism suggests that firms evolve in the same direction in response to outside forces.[287] Mandatory regulatory requirements are usually categorized within the frame of coercive isomorphism.[288] Mimetic isomorphism occurs when firms decide to model their policies and responses on peer firms.[289] When organizations innovate, there comes a point where that innovation catches on, and a firm can gain legitimacy within its field by adopting it.[290] Normative isomorphism occurs where policies, culture, and other personnel decisions are guided by professional norms, and as a result firms end up employing people with shared training and professional ethos.[291] The theory holds that none of these processes are dependent on them being the right move from the standpoint of rational organizational goals; rather, they are processes that simply operate by dint of how firms and industries are organized and socialized.[292] In fact, institutional isomorphism can sometimes operate in opposition to efficiency goals.[293]

Edelman's analysis of legal endogeneity suggests that the conventional understanding of legal compliance as a form *coercive* isomorphism might be incorrect; rather, the spread of symbolic structures takes the form of mimetic or normative isomorphism.[294] Writing with Mark Suchman, Edelman argues that:

> Given ambiguity and complexity in the law, environment-level dynamics such as mimetic and normative isomorphism play a central role in transforming vague legal strictures into concrete organizational practices . . . . [T]he definition of compliance emerges collectively and often cooperatively within an organizational community, and compliant behavior is

285. DiMaggio & Powell, *supra* note 284, at 150.
286. *Id.* at 150–51.
287. *Id.*
288. *Id.*
289. *Id.* at 151–52.
290. *Id.* at 148.
291. *Id.* at 152.
292. *Id.* at 153.
293. *Id.* at 153–54.
294. *See* EDELMAN, *supra* note 235, at 32.

> motivated more by cultural norms and accounts than by the imminent threat of legal sanctions.[295]

Waldman observed isomorphism at work as privacy professionals shared experiences and protocols at industry conferences, which he argues leads to similarities in privacy policies and other symbolic structures across companies.[296] Both Edelman and Waldman appear to believe that these processes of isomorphism lead to a leveling down of substantive compliance, where firms figure out ways to effectively cheat the requirements, and then share that knowledge with others. Indeed, that seems like a likely outcome, but nothing *necessitates* that outcome in every instance.

If industry leaders establish a baseline of substantive compliance with AIA regulation or even beyond-compliance behaviors, both mimetic and normative isomorphism suggest the possibility that other firms might follow suit.[297] From a profit perspective, if industry leaders demonstrate that compliance comes with benefits for the company, either in competitive posture, reputational goodwill, or good relationships with regulators, then mimetic isomorphism suggests that other companies will follow rather than attempt to invent new ways to cheat the system.[298] Normative isomorphism can occur if the industry leaders set the tone for algorithmic systems such that there is a social or professional cost to firms or individual employees for failing to comply, or seeming not to care. The existence of an industry organization such as the Institute of Electrical and Electronics Engineers ("IEEE")[299] or Partnership on AI ("PAI"),[300] which includes among its members most of the largest AI companies[301] and has as a substantial portion of its mission the reduction of algorithmic harms,[302] surely increases the chance that industry leaders can set norms in a way that results in dragging the entire industry toward better approaches over time.

---

295. Mark C. Suchman & Lauren B. Edelman, *Legal Rational Myths: The New Institutionalism and the Law and Society Tradition*, 21 L. & SOC. INQUIRY 903, 922–23 (1996).

296. Waldman, *supra* note 148, at 808.

297. *Cf.* Anthony et al., *supra* note 272, at 112–13 (finding that mimetic, but not coercive, isomorphism can lead to voluntary compliance in the healthcare industry).

298. *See* Jens Beckert, *Institutional Isomorphism Revisited: Convergence and Divergence in Institutional Change*, 28 SOC. THEORY 150, 155 ("Isomorphic institutional change occurs if institutional models exist that institutional entrepreneurs actively seek to imitate because they are interpreted as attractive institutional solutions to the problems being faced.").

299. *Mission & Vision*, IEEE, https://www.ieee.org/about/vision-mission.html [https://perma.cc/VV67-MAB3].

300. P'SHIP ON AI, https://www.partnershiponai.org/ [https://perma.cc/GK48-EGXE].

301. *Our Partners*, P'SHIP ON AI, https://partnershiponai.org/partners/ [https://perma.cc/CV7K-UQU4].

302. *About Us*, P'SHIP ON AI, https://partnershiponai.org/about/ [https://perma.cc/VY62-ZTLV].

A lesson for AIA regulation is that there may be divergence in the ability to effectively implement the regulation's different purposes. One can examine the effects of legal requirements on both an individual firm level and at field-wide scale, and here it may be useful to think about them separately. There are certainly expressive elements to anti-discrimination law,[303] and privacy is worth protecting on a society-wide level for separate reasons from individual ones.[304] However, anti-discrimination and privacy law as they exist are primarily focused on vindicating individual rights. For such laws, big picture efficacy is judged by the sum of individual cases, and the law mainly fails or succeeds by the percentage of cases of harm that it prevents or rectifies. But while AIAs will operate on the level of individual firms, there are no particular individuals' rights to vindicate with an AIA, so the focus can be different. Yes, it would be good to ensure that every firm individually fixes its algorithms, but the purposes of the regulatory scheme are broader. AIA regulation will seek to have firms experiment, think through the problem, and report back in the form of an AIA. On an individual firm level, the thought is that the forced reflexivity will lead to an improvement in outcomes. But the nature of experimentation is that, even presuming good faith and full cooperation, not all efforts will work. The broader goals of AIAs — to increase our understanding of the process, to educate broader society, and to reform the field's culture — may still succeed even if individual firms fail to improve, as long as we learn from it.[305]

The AIA regulation's goals can thus be seen as layered: both short-term and focused on individual firms, and long-term and focused on the industry as a whole. The individual-firm goals would ideally begin as soon as possible to get firms to become more reflexive about their own product design, and also to generate information and documentation that can inform public debate. These regulatory goals are likely to lead to widely varied levels of compliance. The efficacy of long-term goals, however, do not turn on full compliance by every company. After a certain threshold of companies who do comply, the addition of similar documentation from similar companies will likely offer diminishing returns as companies encounter similar troubles with their models and engage similar approaches to fixing them. And from an industry reform standpoint, isomorphism

303. *See, e.g.*, Deborah Hellman, *The Expressive Dimension of Equal Protection*, 85 MINN. L. REV. 1, 14 (2000); Cass R. Sunstein, *On the Expressive Function of Law*, 144 U. PA. L. REV. 2021, 2051 (1996).

304. *See, e.g.*, Robert C. Post, *The Social Foundations of Privacy: Community and Self in the Common Law Tort*, 77 CALIF. L. REV. 957, 964 (1989); HELEN NISSENBAUM, PRIVACY IN CONTEXT 129–50 (2009) (conceptualizing privacy as a respect for the informational norms integral to social contexts).

305. *See* Freeman, *supra* note 178, at 31 (explaining that the experimentalist ethic means that errors are not failures).

suggests that norm-setting by policymakers and industry leaders can move the field over time even where individual firms resist.

There is a degree of hopefulness in this prescription that may not be warranted. Frankly, it is hard to tell how common beyond-compliance behaviors are, and a healthy dose of skepticism is proper. Lest this discussion seem pollyannish in the face of the profit motive, then, I would just note the modesty of the claim. AIAs cannot solve all algorithmic harms, but they can put the industry and regulators in better positions to avoid the harms in the first place and to act on them once we know more. AIA regulation simply would not have the same goals as command-and-control regulation, and thus the point that such regulation can be defeated by institutional logics is potentially less of a problem.[306]

* * * * *

This Part has argued for three main takeaways for legislators that come from recognizing the challenges of AIA implementation on the ground. First is that humility and flexibility are key; the point of creating AIAs is that we don't know what we don't know, and the goals of the AIA will be best accomplished by partnership between regulator and industry, rather than an adversarial approach. Second, there is a need to identify in a concrete way both the substantive values at stake and the minimum standards needed for compliance. Ambiguity in the AIA requirements will likely lead to undermining of the policy within individual firms, but given the lack of existing knowledge and expertise of outside firms, fully prescriptive regulation is not realistic. Thus, directing the policy goals at a high level and deferring the implementation is the natural compromise. Third, legislators should be aiming to capitalize on beyond-compliance behaviors to the extent possible and should be less concerned with policing individual-firm compliance. This puts the emphasis on encouraging industry leaders on social issues and norm development. It is important to recognize that this may be a rocky path. A company could, for example, position itself as an industry leader on AI ethics, and then turn around and fire

306. In fact, insofar as an additional goal of this regulation is norm development and long-term reform of the tech industry — which perhaps it should be — then premature prescriptive regulation could be actively harmful, rather than merely ineffective. *See* Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 680 (2020) ("One set of discrete administrative corrections — prescriptive regulation — is likely to work only in zones where the goal is to control a bad outcome, and where that risk is so acute that it is acceptable if sector-specific control comes at the cost of cross-sectoral principles and norm development."); BAMBERGER & MULLIGAN, *supra* note 236, at 245 ("The greater regulatory control is vested in centralized authorities . . . the more likely corporations are to adopt a compliance-only outlook, which is another way of saying they will narrowly seek to avoid fines and punishments rather than attend to regulatory aims.").

the leadership of their ethics team over the course of three months on a thin pretext.[307] Nonetheless, persistent norm creation by regulators and willing industry leaders could lead to positive outcomes in the longer term.

V. LEARNING FROM THE FIELD

While Part IV argued that full compliance is not necessary to achieve both of the AIA's goals, the more that companies voluntarily substantively comply, the better. Reform within organizations is more likely to be adopted and resistance reduced when the reform closely aligns with the processes and organizational culture of the firms it is meant to operate in.[308] If the goal is to encourage participation and reduce friction, then it is imperative that regulations be written in a way that is legible to the industry. Any implementation of AIAs in law must learn from the technical field itself.

To oversee the AIA process, regulators also need to understand the way that the technology firms and programmers make decisions in practice. As a general rule, firms' adherence to procedures and written rules is more myth than fact,[309] and the extent to which decisions are planned or ad hoc will have implications for the efficacy of AIAs and the extent to which AIAs reflect actual practice. Regulators also need to understand the universe of choices that industry actors themselves see as available. Restrictions may exist because of cost or practicality, or simply lack of imagination. Regulators will not be able to contextualize the limits of the AIAs, or whether they are adequately completed, unless there is a general understanding about what choices are available, and what technology can and cannot realistically accomplish.

As it turns out, there is a growing literature — both in computer science academia and the technology industry — that can point regulators in the right direction. There is a developing set of empirical scholarship, models for documentation, benchmarking standards, and ethical codes, all of which can shine a light on how to reduce barriers and increase substantive compliance within the industry. This Part will demonstrate how each can be helpful for the eventual AIA project.

---

307. Tom Simonite, *A Second AI Researcher Says She Was Fired by Google*, WIRED (Feb. 19, 2021, 7:21 PM), https://www.wired.com/story/second-ai-researcher-says-fired-google/ [https://perma.cc/E97H-LEYD].

308. *See, e.g.*, PRAKASH, *supra* note 279, at 153; Madaio et al., *supra* note 266, at 7 ("[P]articipants . . . felt strongly that AI fairness checklists must be aligned with teams' existing workflows.").

309. *See* Meyer & Rowan, *supra* note 284, at 346.

It is worth underlining that while learning from the industry is important, regulators must be mindful not to cede the role of law and policy to industry. If policymakers were to simply adopt industry best practices into law and call it a day, they would have essentially blessed self-regulation — and there is no reason to think that will be good enough. The role of law and policy is to set the normative concerns and ultimately the safeguards that industry actors must satisfy.[310] "Essentially contested" concepts, such as privacy,[311] fairness,[312] and equity, belong in the political realm,[313] and thus law and policy should dictate what counts as a cognizable impact.[314] What industry can teach regulators are the ways that their production processes interact with those goals, and what steps can be taken to make achieving the goals as painless and costless as possible.

We should also be careful to not attribute greater expertise to private industry than it may have. Lawyers are often wary of math, science, and technology[315] and this manifests as excess deference to the apparent "magic" of it.[316] But while technology companies are very good at building technology and selling it for profit — their core competency — policy should not be designed assuming industry actors possess skills they do not. Many scholars have pointed out the limitations of computer scientists' methodological training in their attempts to fix algorithmic injustice.[317] Some of the challenges that firms will face can be explained by the totalizing force of market logics,[318] but some of it comes down to engineers and product managers not being trained in law and policy, applied ethics, empirical social

---

310. *See* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 407–10 (explaining why it is important to think of "policy" instead of "ethics" or "governance").

311. Deirdre K. Mulligan, Colin Koopman & Nick Doty, *Privacy Is an Essentially Contested Concept: A Multi-Dimensional Analytic for Mapping Privacy*, 374 PHIL. TRANSACTIONS ROYAL SOC'Y A 1 (2016); *see also* NISSENBAUM, *supra* note 304, at 6–11 (situating privacy at the nexus of constantly contested and competing interests); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 486 (2006) (arguing that privacy can only be understood as a set of concepts related by Wittgensteinian "family resemblance").

312. *See* Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 375, 375 (2021).

313. *See* Selbst et al., *supra* note 3, at 61–62 (arguing that contestability is a core part of social concepts such as fairness).

314. *See* Metcalf et al., *supra* note 270, at 736.

315. *See* Jackson v. Pollion, 733 F.3d 786, 788 (7th Cir. 2013) (citation omitted) ("The discomfort of the legal profession, including the judiciary, with science and technology is not a new phenomenon. Innumerable are the lawyers who explain that they picked law over a technical field because they have a 'math block' — 'law students as a group, seem peculiarly averse to math and science.'").

316. Madeleine Clare Elish & danah boyd, *Situating Methods in the Magic of Big Data and AI*, 85 COMMC'N MONOGRAPHS 57 (2018).

317. *See* Barabas et al., *supra* note 15, at 168; Greene et al., *supra* note 263, at 2123; *cf.* Selbst et al., *supra* note 3, at 59 (discussing the limitations of abstraction as a method).

318. *See supra* Section IV.B.

science, or community organizing. These capabilities all require specialized expertise, and telling firms to just think about it harder will not be adequate.

A. Starting with the Technology

The first step for legislators and regulators is to understand the technology itself, and how its various parts implicate legal and normative values. Paul Ohm and David Lehr have provided a guide for lawyers to the steps in designing a machine learning system.[319] They describe eight steps to the machine learning process: problem definition, data collection, data cleaning, summary statistics review, data partitioning, model selection, model training, and model deployment.[320] They argue that legal scholarship has been attentive to the early stages of the machine learning process, but far less attentive to the middle stages of the process, attention to which can indicate more points of intervention and types of possible intervention.[321] Not every engineer would separate the steps out precisely this way,[322] but the specific breakdown is not so important. Regulators who work with companies will be better equipped to understand their particular processes if they understand a more complete generic model such as this.

Lehr and Ohm's treatment of calls for explainability — which they call reason-giving — illustrates why a level of technical detail is important to understand.[323] They note that legal scholars frequently refer to the algorithmic system as a black box, obscuring the ways that it can and cannot actually be interrogated.[324] Lehr and Ohm distinguish between seeking varieties of reason-giving that are illogical because they misunderstand the technology (asking which feature "caused" a result or why the algorithm offered a certain result as a general matter) and those that are useful (how much different input variables or changes in input variables affect the outcome).[325] If AIA regulation were to demand the former type of explanation, it would leave engineers shaking their heads, not only unable to answer, but convinced the entire AIA project is a waste of time.

---

319. *See generally* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017).

320. *Id.* at 669–70.

321. *Id.* at 715.

322. *See* Harini Suresh & John Guttag, *A Framework for Understanding Unintended Sources of Harm Throughout the Machine Learning Life Cycle*, ARXIV, June 15, 2021, at 1, 2, https://arxiv.org/pdf/1901.10002v4.pdf [https://perma.cc/YN3L-GRPH] (describing the machine learning process in six stages).

323. Lehr & Ohm, *supra* note 319, at 705–10.

324. *Id*. at 706.

325. *Id.* at 708–10.

Harini Suresh and John Guttag make similar points.[326] They note that people often speak about "bias" in machine learning systems while failing to distinguish between effects of bias that can come from the different stages of the machine learning process, including "historical bias" that is prior to development, and "deployment bias" that is related to how the system is used, and therefore subsequent to it.[327]

These scholars demonstrate a basic understanding of technology that regulators will need, but that is still not enough for the AIA for three reasons. First, theirs are pictures of the technology,[328] not the institutional processes that create the technology. Both are necessary. The institutional process, not the technology, is what will determine the success of the intervention, and the technology that a firm ends up producing will be inevitably shaped by the specific processes that create it.[329] Second, studying the attempts to make the technology more accountable and equitable is as important as studying its development. People in these firms are already working on this and we need to understand what works and what doesn't. Third, we need to understand how the choices made play out in practice, both for the purposes of ongoing monitoring, and to understand how they affect communities and how companies are capturing that information.[330]

B. Looking to Qualitative Empirical Research

Empirical research into the technology firms themselves will also be essential in determining their work culture and how they make decisions.[331] Several examples have been published in just the last few

326. *See* Suresh & Guttag, *supra* note 322, at 5–9.

327. *Id.*

328. The picture from Suresh and Guttag's is more expansive, in that it includes the effects of society outside the firm's walls, *see generally id.*, but does not address organizational issues with respect to product development.

329. *See generally* TREVOR J. PINCH & WIEBE E. BIJKER, *The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other*, *in* THE SOCIAL CONSTRUCTION OF TECHNOLOGICAL SYSTEMS: NEW DIRECTIONS IN THE SOCIOLOGY AND HISTORY OF TECHNOLOGY 11, 21–33 (Wiebe E. Bijker et al. eds., 1987) (arguing that the final form of technology ends up shaped by competing interest groups); Wanda J. Orlikowski & Stephen R. Barley, *Technology and Institutions: What Can Research on Information Technology and Research on Organizations Learn from Each Other?*, 25 MGMT. INFO. SYS. Q. 145, 145–46 (2001) (arguing that technology is shaped by organizational processes); SHEILA JASANOFF, *Ordering Knowledge, Ordering Society*, *in* STATES OF KNOWLEDGE 13, 15–43 (Sheila Jasanoff ed., 2004) (arguing that "co-production" is the proper framework to understand the relationship between society and science or technology).

330. Kate Crawford & Ryan Calo, *There Is a Blind Spot in AI Research*, 538 NATURE 311, 313 (2016) (arguing that we suffer from a lack of "social systems analysis" of data systems already in use).

331. *E.g.*, Samir Passi & Steven J. Jackson, *Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects*, PROC. ACM ON HUM.-COMPUT. INTERACTION, Nov. 2018, at 1, 2 ("Contemporary understanding of data science in research is largely based on the analysis of data science work in academic and research

years that demonstrate how useful such research can be. Samir Passi and Solon Barocas's ethnographic study tracing "problem formulation in practice"[332] is a great example. The paper draws on a six-month ethnographic study with a data science team at an automotive financing lead generation firm.[333] Passi and Barocas detail the highly collaborative and iterative process that goes into problem formulation — the very first step of building this technology.[334] While building its system, the firm repeatedly encountered concepts that were too difficult to define or directly measure with the data they had.[335] Passi and Barocas charted how the goals of the project changed along the way as a result.[336] The company started out with an unstated objective to minimize churn rate (objective #1) — that is, retain more of the dealer-customers — and figured that they could accomplish that by improving the quality of the leads they offered (objective #2), which then had to be defined as finding the best leads for each dealer (objective #3), changing the task to a matching problem.[337] This led to the idea of "financeability" as a criterion, which they noted was related to predicting credit scores of leads (objective #4).[338] But it turned out they only had data on 50-point credit score ranges, from different sources that used different overlapping ranges. So, to simplify, they aimed to predict whether a lead's credit score was above or below 500 (objective #5).[339]

This study perfectly embodies the kinds of decisions that should be laid out in an AIA. The company started with a goal of minimizing churn, and then for a host of practical reasons, ended up predicting whether a lead was above or below a credit score. Irrespective of whether these specific decisions should be held to be defensible or not, they are exactly the information that an AIA would seek: how companies understand the problems they are solving, the kinds of constraints that companies face, and how they go about solving the

---

sites; shaped by limits of access, confidentiality, and non-disclosure, the large body of applied data science work in corporate settings has received much less attention."); Ben Hutchinson et al., *Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 560, 560 (2021) (developing an accountability framework based on a deep understanding of engineering practices); Katie Shilton, *Values Levers: Building Ethics into Design*, 38 SCI. TECH. & HUM. VALUES 374, 375 (2013) (using ethnography to understand how certain practices and aspects of the workplace environment affect engineers' ability to importing social values into design).

332. Samir Passi & Solon Barocas, *Problem Formulation and Fairness*, PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 39, 41 (2019).

333. *See id.* at 43.

334. *Id.* at 43–44.

335. *Id.* at 44.

336. *Id.* at 44–46.

337. *Id.* at 46.

338. *Id*.

339. *Id.*

problems. AIAs should be designed so that we can reconstruct this story.

Another example comes from Mark Sendak and colleagues, a team of researchers studying the implementation of Duke University Hospital's Sepsis Watch program.[340] Sepsis is a dangerous and rapid-onset condition in hospitals.[341] Yet not all doctors agree on how to detect or diagnose sepsis, so the Sepsis Watch team was tasked with creating an early-response warning system.[342] Their article explains many of the choices the team made. For example, the Sepsis Watch team deprioritized explainability.[343] Because speed was important, explainability tools were not that effective, and because hospitals already have experience with highly distributed tasks where not every person understands each part, the team felt that explainability was worth trading in for more speed.[344] The team also explained the decision to set the system up such that a rapid response nurse was the primary user, who would not diagnose sepsis, but would screen alerts to determine when to escalate.[345] This decision was made with clinician input because of history with a prior sepsis detection tool triggering too often, resulting in alert fatigue.[346] The team also learned that while there was a tendency to defer to the tool in the beginning, medical professionals became more comfortable over time, and that despite a lack of explainability, "new sets of expertise . . . emerge[d] and in fact enhanced the use of the machine learning driven tool."[347]

Like Passi and Barocas's research, Sendak and colleagues' paper illustrates how important decisions were made and contextualizes them. The paper explains consequential decisions and argues for why they were correct within the context in which they were operating. It makes explicit appeals to balancing normative concerns where trustworthiness and speed were prioritized over explainability. It also demonstrates the effects of not only domain awareness, but also the history of the specific institution. Sendak and colleagues studied the effects of the system as it was running and how interactions with it changed over time. Work like this can show regulators the sorts of things to look for in an AIA of the integration of new AI technology into a different context.

---

340. Sendak et al., *supra* note 94, at 99.
341. *Id.* at 102.
342. *Id.*
343. *Id.*
344. *Id.*
345. *Id.*
346. *Id.*; *see also* Michael Greenberg & M. Susan Ridgely, *Clinical Decision Support and Malpractice Risk*, 306 J. AM. MED. ASS'N 90, 90 (2011) (discussing alert fatigue in the medical field).
347. Sendak et al., *supra* note 94, at 106.

Michael Madaio and colleagues' research on the potential use of checklists in technology firms may be the most useful example. They implemented a co-design process to create a checklist, involving practitioners from companies already working on AI ethics and fairness.[348] They conducted workshops and interviews about the process and how AI ethics efforts work in practice at the participants' firms.[349] Their work has several notable pieces of information for AIA regulation. First, it provides something close to a model AIA prompt. Though the word "checklist" tends to evoke closed-ended — and gameable — questions,[350] their checklist does not, instead including items that invite open-ended responses, such as:

> Envision system and its role in society, considering:
> - System purpose, including key objectives and intended uses or applications . . .
> - Sensitive, premature, dual, or adversarial uses or applications . . .
> - Expected deployment contexts (e.g., geographic regions, time periods)
> - Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
> - Expected benefits for each stakeholder group, including demographic groups
> - Relevant regulations, standards, guidelines, policies, etc.[351]

Their checklist also includes elements at different stages of product development.[352] This means that a checklist like this, co-designed with practitioners and open-ended, is almost a recipe for AIA generation, and a potential model that the regulation could even adopt, where the answers would become the AIA.

---

348. Madaio et al., *supra* note 266, at 4–5.

349. *Id.* at 1.

350. Indeed, some participants in their study worried about exactly such a type of quantified checklist because it can be so easily undermined. *Id.* at 8. The study authors acknowledged that the checklist framing might be problematic, and argued that it should be "designed to prompt discussion and reflection that might otherwise not take place." *Id.* at 10.

351. *Id.* at 6.

352. *Id.*

Their interviews are also very informative as to how firms think about ethics and fairness and would respond to AIAs. They confirmed managerialist tendencies, finding that "organizational culture typically prioritizes 'moving fast' and shipping products over pausing to consider fairness . . . ."[353] They also found, as in other contexts, that having leadership embrace ethics matters.[354] For new information, they found a practical desire for flexibility in the checklists.[355] When presented with a generic checklist, the participants noted that each would have to customize it to their firms or even product teams within firms, leading participants to prefer a principles-based approach that product teams could implement.[356]

One of Madaio and colleagues' most notable observations accords with what has been seen in other industries: the voluntary use of checklists increases when reporting comes at points in the process that make it less burdensome.[357] These points are sometimes called "pause points."[358] In today's technology industry, where algorithmic models are often finished with a cycle of tweaks and tests until it feels right,[359] interrupting the process to document everything would engender strong resistance, and thus understanding pause points is important.

Unfortunately, Madaio and colleagues also confirmed that there is much work to be done on engaging affected communities.[360] They note that companies have tools for "user experience" testing, but that such testing is about the user, not affected parties.[361] "For example, a UX researcher working on a predictive policing system might solicit feedback from the police — i.e., the intended users of the system — but fail to engage with the communities most likely to be affected by

353. *Id.* at 10.

354. *See* Madaio et al., *supra* note 266, at 9.

355. *Id.* at 9–10.

356. *See id.* at 8–9.

357. *See id.* at 7. *See generally* Barbara K. Burian, *Design Guidance for Emergency and Abnormal Checklists in Aviation*, 50 HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING PROC. 106 (2006) (describing such points in aviation checklists); Barbara K. Burian, Anna Clebone, Key Dismukes & Keith J. Ruskin, *More than a Tick Box: Medical Checklist Development, Design, and Use*, 126 ANESTHESIA & ANALGESIA 223 (2018) (in anesthesia checklists); Asaf Degani & Earl L. Wiener, *Cockpit Checklists: Concepts, Design, and Use*, 35 HUM. FACTORS 345 (1993) (in another example of aviation checklists); Brigette M. Hales & Peter J. Pronovost, *The Checklist: A Tool for Error Management and Performance Improvement*, 21 J. CRITICAL CARE 231 (2006) (in medical and critical care checklists); PRAKASH, *supra* note 279, at 153 (in environmental policy processes within firms).

358. Madaio et al., *supra* note 266, at 7.

359. *See* Seda Gürses & Joris van Hoboken, *Privacy After the Agile Turn*, *in* THE CAMBRIDGE HANDBOOK OF CONSUMER PRIVACY 579, 582–83 (Evan Selinger, Jules Polonetsky & Omer Tene eds., 2018) (describing the software industry's shift to "agile" development, marked by, among other things, "short development cycles," "continuous testing," and a preference for "working software over comprehensive documentation").

360. Madaio et al., *supra* note 266, at 10.

361. *Id.*

the system's use."[362] This can be generalized; there is a robust literature on human-computer interaction, but it is primarily about the users of systems, with little about the effects on the people who are subject to decisions that the systems are used for.[363] More work is needed there.

There are several other examples of new empirical research,[364] but hopefully these suffice to illustrate the variety of important insights that will come from studying the organizations that produce AI technologies. Given the surging interest in these issues, more empirical work is likely underway. But qualitative empirical research takes time and resources. In the meantime, other sources can be consulted to understand the pause points and decisions in the process.

C. Documentation and Testing Standards

Part of the challenge facing regulation of AI systems is a lack of documentation and testing standards.[365] When information is presented in standard form, it is easier to digest and incorporate into systems.

---

362. *Id.*

363. Computer science researchers working on explanation and interpretability are trying to fill this gap by creating frameworks to identify diverse stakeholders. *See, e.g.*, Harini Suresh, Steven R. Gomez, Kevin K. Nam & Arvind Satyanarayan, *Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., May 2021, at 1, https://arxiv.org/pdf/2101.09824.pdf [https://perma.cc/W2YM-VCM2]; Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett & Supriyo Chakraborty, *Stakeholders in Explainable AI*, 2018 A.I. IN GOV'T & PUB. SECTOR, https://arxiv.org/pdf/1810.00184.pdf [https://perma.cc/Y9SG-63YD]; Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece & Supriyo Chakraborty, *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems*, ICML WORKSHOP ON HUM. INTERPRETABILITY IN MACH. LEARNING 8, 8 (2018); Foad Hamidi, Morgan Klaus Scheuerman & Stacy M. Branham, *Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., Apr. 2018, at 1; Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler & Jeff Warshaw, *A Qualitative Exploration of Perceptions of Algorithmic Fairness*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., Apr. 2018, at 1, 1.

364. *See* Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík & Hanna Wallach, *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., May 2019, at 1 (studying private sector ML practitioners who are starting to do fairness work); Michael Veale, Max Van Kleek & Reuben Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., Apr. 2018, at 1 (studying machine learning systems in the public sector in five OECD countries); ALEXANDRA MATEESCU & MADELEINE CLARE ELISH, AI IN CONTEXT 4 (2019) (studying automation in grocery retail and farm management); Metcalf et al., *supra* note 258, at 455 (studying "ethics owners" at a variety of firms that create data-centric technologies).

365. Ben Hutchinson et al., *supra* note 331, at 6 ("The information that is shared as a necessary (but not sufficient) precondition for accountability is referred to technically as *accounts*. The recording of dataset accounts is at its most fundamental a question of bookkeeping, but the details are critical: which books should be kept, what are their stories, and who are their authors?").

Standards also make evaluation easier, as the information is structured for oversight, and a regulator can ensure that each element of the documentation standard is met. There is also a downside to documentation standards, however. The structure of standard reporting requirements will structure the thinking around them — the kinds of testing that firms do, how they evaluate the impacts of their systems, and which impacts they think to evaluate — and this structured thinking ends up becoming hard to see once a standard is settled and in place.[366] This downside is why it is important that AIAs remain open-ended, especially as regulators begin to learn what questions are important to even ask. This tradeoff between the need to categorize information and the loss of flexibility is inevitable and exists in every information system — but is only truly visible while the standards are unsettled.[367]

Despite tradeoffs, standards are important for modern industry. They allow comparison, interchangeability, evaluation, and reliability. AI is no exception, but it is a young field, so uniform standards are lacking. Various actors in industry and academia have in the last few years proposed some documentation and testing standards, and more are in the process of development. AIA regulation can learn a lot from these proposed standards.

The first approach to documentation is to label datasets.[368] Many of the problems with AI can be attributed to the use of datasets in inappropriate contexts — programmers often train models with canonical datasets that can have problematic aspects,[369] or they purchase data off the shelf that has no guarantees of generalizability or validity in a different context.[370] The idea is that people who create a dataset should explain its characteristics and limitations — for example, the purpose and context for which it was created,[371] what pre-processing

366. *See generally* Geoffrey C. Bowker & Susan Leigh Star, *Invisible Mediators of Action: Classification and the Ubiquity of Standards*, 7 MIND, CULTURE & ACTIVITY 147 (2000) (discussing the ubiquity, invisibility, and influence of standards).

367. *See* GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT 33, 38 (1999) (noting that classifications and standards "disappear almost by definition" because we are "quite schooled at ignoring both").

368. Timnit Gebru et al., *Datasheets for Datasets*, COMMC'NS ACM, Dec. 2021, at 86, 86; Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph & Kasia Chmielinski, *The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards*, *in* DATA PROT. & PRIV. 4 (Dara Hallinan et al. eds., 2020); Emily M. Bender & Batya Friedman, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 TRANSACTIONS ASS'N FOR COMPUTATIONAL LINGUISTICS 587, 588 (2018).

369. *See* Abebe Birhane & Vinay Prabhu, *Large Image Datasets: A Pyrrhic Win for Computer Vision?*, 2021 IEEE/CVF WINTER CONF. ON APPLICATIONS COMPUT. VISION 1, 2.

370. Bender & Friedman, *supra* note 368, at 587–88.

371. Gebru et al., *supra* note 368, at 88.

or data cleaning was done,[372] or the demographics represented in the data.[373] This information would allow data practitioners to choose their datasets efficiently and wisely, thus improving the quality of their AI. If datasets are labeled and someone purchases a hiring algorithm, they can potentially choose between models trained in different industries or on groups of people with different demographics than your applicant pools, to get as close to your own as possible.

Versions of this dataset labeling approach were proposed by three different research teams. Each chose a different metaphor for the documentation, but their conclusions were similar. Timnit Gebru and colleagues draw on the idea of a "datasheet" from the electronics industry.[374] Sarah Holland and colleagues instead rely on the model of a nutrition label.[375] Emily Bender and Batya Friedman, working in the natural language processing context, propose "data statements," which are not modeled on a well-known system, but are similarly intended to be compact, yet detailed.[376] Each of these informational structures is designed to deliver critical information in an efficient and recognizable way, allowing for a more considered use of datasets with little effort once a standard is agreed to.[377]

Other papers offer analogues for later stages of the process: "Model Cards" for pre-trained models[378] or "FactSheets" for final AI services.[379] Margaret Mitchell and colleagues (including Gebru) directly built on the datasheets model, noting that "[w]here Datasheets highlight characteristics of the data feeding into the model, we focus on trained model characteristics such as the type of model, intended use cases, information about attributes for which model performance may vary, and measures of model performance."[380] The premise is the

372. *Id.* at 89–90.

373. Holland et al., *supra* note 368, at 17–18.

374. Gebru et al., *supra* note 368, at 86. In that industry, any circuit element you buy off the shelf comes with a datasheet that includes: a general description of the part; its typical characteristics; its performance in different environments such as hot or cold temperatures, noisy environments, or high and low power supplies; and the points at which it breaks. *See, e.g.*, ANALOG DEVICES, *Datasheet for AD741, Low Cost, High Accuracy IC Op Amps*, https://analog.com/media/en/technical-documentation/data-sheets/AD741.pdf [https://perma.cc/8L8B-G7MB].

375. Holland et al., *supra* note 368, at 1.

376. Bender & Friedman, *supra* note 368, at 587.

377. *See* Holland et al., *supra* note 368, at 23 (arguing that the proposal will offer "data specialists . . . a better, more efficient process of data interrogation").

378. Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 220, 220 (2019).

379. Matthew Arnold et al., *FactSheets: Increasing Trust in AI Services Through Supplier's Declarations of Conformity*, 63 IBM J. RSCH. DEV. 6:1, 6:1 (2019).

380. Mitchell et al., *supra* note 378, at 221 ("Model cards provide a way to inform users about what machine learning systems can and cannot do, the types of errors they make, and additional steps that could be taken to create more fair and inclusive outcomes with the technology."). Mitchell and Gebru developed the framework while leading Google's Responsible AI division before being controversially fired. *See supra* note 307 and accompa-

same, just at one level of abstraction higher. A model card, according to Mitchell and colleagues, should include "quantitative evaluation results to be broken down by individual cultural, demographic, or phenotypic groups, domain-relevant conditions, and intersectional analysis combining two (or more) groups and conditions" as well as "the motivation behind chosen performance metrics, group definitions, and other relevant factors."[381] Matthew Arnold and colleagues operate one level of abstraction higher still, proposing FactSheets for full AI services, and including questions about the purposes, internal algorithms used, testing of internal components for bias, and any remediation steps.[382] These papers all cite each other, with the idea being that the documentation approaches of the layers are complementary, and should all be included simultaneously.[383] Finally, Kacper Sokol and Peter Flach have proposed a regime of "explainability fact sheets" to get at similar kinds of questions, but targeting the explanation layer, as discussed in the loan denial example.[384]

These interventions are useful for AIA regulation for a few reasons. Developers see their work product as existing in different stages, and these calls for documentation tell regulators what the legible, discrete units are in the product pipeline. How the different research teams broke up the product pipeline to argue for different intervention points is itself useful information about the commonly understood stages of the development workflow. Holland and colleagues' paper even helpfully includes an entire diagram of the data pipeline to explain how the nutrition label would affect downstream results.[385] Looking to this work suggests that the initial focus, at least, should be concerned with data provenance, model testing, and evaluating top level AI services.

A second way in which documentation interventions are helpful is to outline different ethical concerns of people working in the industry, focused on accountability. The deliverables that each of these interventions ask for are likely to be legible to all engineers working in accountability, and perhaps represent concerns that engineers already think to test for. This provides a good starting point for a regulator to see what kinds of information they should be seeing in an AIA, but

---

nying text. Google has been trying to use Model Cards in practice since. *See Model Cards*, GOOGLE, https://modelcards.withgoogle.com/about [https://perma.cc/H2PK-HMVH].

381. Mitchell et al., *supra* note 378, at 221.

382. Arnold et al., *supra* note 379, at 6:10.

383. *See* Mitchell et al., *supra* note 378, at 221 ("Each model card could be accompanied with Datasheets, Nutrition Labels, Data Statements, or Factsheets, describing datasets that the model was trained and evaluated on."); Gebru et al., *supra* note 368, at 91–92 (discussing model cards and fact sheets).

384. Sokol & Flach, *supra* note 13, at 1.

385. Holland et al., *supra* note 368, at 2.

also to scrutinize what kinds of impacts people in the industry might regularly miss that may need a more direct push.[386]

The third way documentation interventions are helpful — assuming any means of labeling becomes an accepted standard — is in the setting of a documentation standard. If datasets are all labeled in the future with datasheets, for example, then when conducting an AIA, the regulator will want to know why a particular dataset was used and whether it was appropriate, and the engineer who performed the AIA would point to the datasheet as justification. The datasheet may or may not turn out to be *good* justification, but its existence structures the conversation between regulator and regulated to make it more likely to be productive.

D. Ethical Frameworks and Social Impact Assessment

The last things regulators could potentially look to in the industry are ethical and self-regulatory frameworks. For example, the IEEE is a respected standards organization in computer science, and it recently issued Standard IEEE 7010, an industry standard for assessing ethical and social impact of AI.[387] The Partnership on AI is a trade group seeking to develop ethical standards for the AI industry.[388]

Outside of the biggest industry groups, "AI ethics" has become almost an industry unto itself, with many different ethical codes proposed by different companies and organizations.[389] Theoretically, analysis of these ethical codes could tell us about the efforts to make technology more internally accountable, or about the ethical principles that technology companies seek to adhere to on their own, but in practice many of these individual sets of ethical principles are seen as "ethics washing" — little more than attempts to ward off regulation with claims to self-regulation.[390] Moreover, as Hirsch and colleagues

386. *See* Metcalf et al., *supra* note 270, at 735 (arguing that what counts as an "impact" within an AIA is co-constructed between stakeholders, and is not necessarily synonymous with "harm").

387. *See* Daniel Schiff, Aladdin Ayesh, Laura Musikanski & John C. Havens, *IEEE 7010: A New Standard for Assessing the Well-Being Implications of Artificial Intelligence*, IEEE INT'L CONF. ON SYS., MAN & CYBERNETICS 1, 1 (2020).

388. *See About ML*, P'SHIP ON AI, https://partnershiponai.org/workstream/about-ml/ [https://perma.cc/P7HY-4BAV].

389. Mittelstadt, *supra* note 145, at 501 (stating that as of 2019, "at least 84 . . . 'AI Ethics' initiatives have published reports describing high-level ethical principles, tenets, values, or other abstract requirements for AI development and deployment").

390. *See* Bietti, *supra* note 258, at 210 ("[T]he term has been used by companies as an acceptable façade that justifies deregulation, self-regulation or market driven governance, and is increasingly identified with technology companies' self-interested adoption of appearances of ethical behavior. We call such growing instrumentalization of ethical language by tech companies 'ethics washing.'"); Birhane & Prabhu, *supra* note 369, at 10 ("We are up against a system that has veritably mastered *ethics shopping*, *ethics bluewashing*, *ethics*

find, some companies do in fact try to be ethical, but even where ethical codes exist they make decisions with informal gut-checks like "Would my mother think this is okay? Would I want this to happen to my kid?"[391] While we need not dismiss ethics' relevance to the conversation in general,[392] the ethical codes that have shown up in the last five years do not seem to offer enough to work with.

Industry ethical codes aren't the only ones to look at, though. A recent study by the IEEE shows that ethical codes promulgated by industry are far narrower than the ones put out by NGOs, which are more robust.[393] Similarly, some ethics statements or calls for social impact assessment come from the academic sector.[394] The lack of incentives for profit or to evade stricter regulations should render these efforts more trustworthy, and thus useful to examine. While industry ethical codes can be an important source of understanding about the industry's priorities, they cannot truly be the source of substantive understanding regarding what counts as harmful impacts. That is a role better suited to affected communities and the political process.

* * * * *

This Part has examined the work of technology practitioners and academics to draw out lessons for future AIA regulations. It argued that there are four sources of knowledge that can be helpful to legislators or regulators. First, a basic understanding of the technology is necessary to have the AIA requirements be legible and taken seriously by the people charged with performing the assessments. Second, regulators will want to understand the organizational processes by which algorithmic solutions are developed, and can look to the growing set of qualitative empirical research on this subject. Third, documentation standards are emerging from the technical industry and academia, which are useful both for what they reveal about how practitioners see the different stages of their work and for the standardization of documentation itself, which AIAs can borrow from to make evaluation simpler on all sides. Fourth, regulators can look to the emerging AI ethics codes and discussion for an understanding of how the industry sees the accountability problems and a view of industry's potential blind spots. While the ultimate normative goals of AIA regulation must remain in the hands of policymakers, rather than being out-

lobbying, *ethics dumping*, and *ethics shirking*.") (describing the failures of ethics in machine vision, in particular).

391. HIRSCH ET AL., *supra* note 238, at 50.

392. *See generally* Bietti, *supra* note 258, at 210.

393. Daniel Schiff, Jason Borenstein, Justin Biddle & Kelly Laas, *AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection*, IEEE TRANSACTIONS ON TECH. & SOC'Y 1, 8–9 (2021).

394. *See, e.g.*, Diakopoulos et al., *supra* note 35.

sourced to industry, there is a great deal to learn from industry in shaping the AIA requirements, which can make them more effective and efficient.

VI. CONCLUSION

This Article has argued that AIA regulations are a limited but central form of public oversight of algorithms. AIAs have two broad goals: leading companies to think earlier about social impacts and head off problems before they start, and educating the public about how decisions are made within firms that build algorithmic models. Though there is much confusion around the term "algorithmic impact assessment," there are certain elements of an impact assessment framework that are vitally important in our current moment: open-ended questions; pre-deployment assessment; oversight, whether through transparency or closed-door meetings with regulators; and community involvement.

But AIA regulations are in a tight spot. They inherently rely on the cooperation of the private sector. This means that regulations must be flexible and invite cooperation, but must also resist the organizational motivations and logics that will tend to undermine regulations if not spelled out strictly enough. Worse, profit-centered enterprises cannot generally be trusted to meaningfully accept regulation or self-regulate in good faith. Technology companies, in particular, have repeatedly promised to center ethical concerns and human rights within their operations, but have repeatedly failed to provide any details about their implementation.[395] Thus any regulation will need to require enforceable minimum substantive standards. Even with that guardrail, there will still be bad faith actors who either try to cut corners or have a business model that relies on a harmful product. We can imagine that a company like Clearview AI that tries to operate in secret, selling facial recognition to police departments and large corporations,[396] knows fully well that its technology is harmful, and figures that squishy moral values are merely preventing its competitors from entering the market. In cases such as these, AIAs may not work on an individual firm level, since collaboration between regulator and regulated does presume a degree of good faith, and AIAs are thus a limited tool.

---

395. *See generally* Amy Brouillette, *Key Findings: Companies Are Improving in Principle, but Failing in Practice*, 2020 RANKING DIGITAL RIGHTS CORPORATE ACCOUNTABILITY INDEX, https://rankingdigitalrights.org/index2020/key-findings [https://perma.cc/U5AY-FRKH].

396. Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Mar. 18, 2021), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [https://perma.cc/Z8BU-PVWA].

Corporate intransigence aside, however, AIAs may still do some good if even some companies willingly comply. Because companies are moved by internal norm entrepreneurs — who occasionally engage in beyond-compliance behavior — and institutions within a field tend to follow industry leaders, the global goals of AIAs to reform the technical industry over time can potentially be achieved even with only partial compliance. The goals of regulation should therefore be to encourage rather than attempt to force compliance.

Finally, one way to encourage substantive compliance and empower industry leaders is to learn from industry and, where possible, to meet the technical field where it is. An AIA requirement must be legible to the firms that will be preforming the assessments so that they can see the value and collaborate usefully, rather than treat the AIA as a series of boxes to check. To that end, the AIA should be designed around an understanding of the technology itself, the culture and structure of organizations that make the technology, and the emerging benchmarks and standards of the technology industry. Law and policy cannot defer to the industry on the ultimate policy goals; rather, those must be defined by the democratic process. But there is a great deal of flexibility in the law's implementation, and learning from the field will ultimately give AIAs a far greater chance of success.