BEYOND THE "BLACK BOX"

CHARLOTTE A. TSCHIDER†

ABSTRACT

As algorithms have become more complex, privacy and ethics scholars have urged artificial intelligence (AI) transparency for purposes of ensuring safety and preventing discrimination. International statutes are increasingly mandating that algorithmic decision-making be explained to affected individuals when such decisions impact an individual's legal rights, and U.S. scholars continue to call for transparency in automated decision-making.

Unfortunately, modern AI technology does not function like traditional, human-designed algorithms. Due to the unavailability of alternative intellectual property (IP) protections and their often dynamically inscrutable status, algorithms created by AI are often protected under trade-secrecy status, which prohibits sharing the details of a trade secret, lest destroy the trade secret. Furthermore, dynamic inscrutability, the true "black box," makes these algorithms secret by definition: even their creators cannot easily explain how they work. When mandated by statute, it may be tremendously difficult, expensive, and undesirable from an IP perspective to require organizations to explain their AI algorithms. Despite this challenge, it may still be possible to satisfy safety and fairness goals by instead focusing on AI system and process disclosure.

This Article first explains how AI differs from historically defined software and computer code. This Article then explores the dominant scholarship calling for opening the black box and the reciprocal pushback from organizations likely to rely on trade secret protection—a natural fit for AI's dynamically inscrutable algorithms. Finally, using a simplified information fiduciary framework, I propose an alternative for promoting disclosure while balancing organizational interests via public AI system disclosure and black-box testing.

TABLE OF CONTENTS

† Assistant Professor, Loyola University Chicago School of Law. I would like to thank the Junior Intellectual Property Scholars Conference participants, the Internet Works in Progress Conference participants, the Law and Technology Workshop participants, and the AALS Intellectual Property Section's members, especially W. Nicholson Price II, Rachel Sachs, Ana Santos Rutschman, Andrew Michaels, and Lucas Osborn, for their helpful suggestions in the development of this Article.

INTRODUCTION

> "Both the man of science and the man of art live always at the edge of mystery, surrounded by it . . . ."
>
> – J. Robert Oppenheimer[1]

AI, often the territory of science fiction, is officially part of the average American's daily life. Most Americans wake up in the morning and check their e-mail or social media sites, which are subject to automated filtration, "smart" replies, and nudging reminders.[2] AI infrastructure used by social media sites and customer support chatbots have changed how, with whom, and with what technologies we interact.[3] AI has long been part of commercial experiences, including product recommendations, search criteria, and ride-share apps.[4] And transportation has similarly been affected: a 2015 survey of airline Boeing 777 pilots revealed that they only spend seven minutes actually flying the aircraft during a typical flight.[5] Although sentient AI might only exist in science fiction for now, AI-enabled automated decisions are being made on a regular basis today.

For more than a decade, privacy advocates have identified the need for greater transparency related to data use and automated algorithmic decision-making to prevent discrimination and buoy fairness, while product-safety advocates have questioned the safety of AI-enabled systems.[6] Algorithms, increasingly automated and not subject to human review, make decisions that affect every aspect of an individual's life, including whether your car stops or accelerates (self-driving cars),[7] whether you will be approved for a loan (automated credit check),[8] whether you receive a second interview (employment screening and background checks),[9] or whether you need more insulin (medical device AI).[10] There

---

1. J. ROBERT OPPENHEIMER, THE OPEN MIND 133–46 (1955), *reprinted in* Fred W. Friendly & Michael G. Wood, Seminar Reports, *J. Robert Oppenheimer: The Man and His Case*, 2 COLUM. DAILY SPECTATOR, no. 4 (Mar. 3, 1975) at 4.

2. *16 Examples of Artificial Intelligence (AI) in Your Everyday Life*, MEDIUM (Sept. 26, 2018), https://medium.com/@the_manifest/16-examples-of-artificial-intelligence-ai-in-your-everyday-life-655b2e6a49de.

3. *Id.*

4. *Id.*

5. *Id.*

6. *See, e.g.*, Katyanna Quach, *Remember the Uber Self-Driving Car That Killed a Woman Crossing the Street? The AI Had No Clue About Jaywalkers*, THE REG. (Nov. 6, 2019, 5:48 AM), https://www.theregister.co.uk/2019/11/06/uber_self_driving_car_death/ (describing how AI caused the death of a pedestrian).

7. *On the Road to Full Autonomy: Self-Driving Cars Will Rely on AI and Innovative Memory*, MICRON, https://www.micron.com/insight/on-the-road-to-full-autonomy-self-driving-cars-will-rely-on-ai-and-innovative-memory (last visited Feb. 23, 2021).

8. Daniel Faggella, *Everyday Examples of Artificial Intelligence and Machine Learning*, EMERJ (Apr. 11, 2020), https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/.

9. Drew Harwell, *A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job*, WASH. POST (Nov. 6, 2019, 10:21 AM),

are countless applications for AI algorithms, some that may impact human rights and freedoms, and others that may pose safety concerns or risk property damage.[11] However, these technologies, in varying degrees, have the potential to solve the most impossible of human problems.[12]

The law has not kept pace with AI technologies; retaining procedural restrictions on data collection and use rather than facilitating safe, responsible, and fair collective social benefit.[13] Privacy and data protection laws, for example, have previously addressed the potential for data overuse and abuse through upfront privacy notices coupled with consent, which fails to facilitate actual, human choice regarding data handling practices.[14]

---

https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/; *AI-Powered Core*, CHECKR, https://checkr.com/platform/foundation/ai-powered (last visited Feb. 23, 2021).

10. Jessica Kent, *Artificial Intelligence, Big Data to Improve Diabetes Management*, HEALTH IT ANALYTICS (Aug. 13, 2019), https://healthitanalytics.com/news/artificial-intelligence-big-data-to-improve-diabetes-management.

11. AI implications for individual rights and freedoms have been examined for some time, initially introduced in relation to automated credit decisions. Buoyed by concerns around autonomous car safety, organized policy groups like the Brookings Institute have advocated for a products liability model to respond to potential physical safety issues introduced by AI. *See* John Villasenor, *Products Liability Law as a Way to Address AI Harms*, BROOKINGS INST. (Oct. 31, 2019), https://www.brookings.edu/research/products-liability-law-as-a-way-to-address-ai-harms/.

12. *See infra* Part I.

13. For example, privacy advocates and, indeed, laws like the Health Insurance Portability and Accountability Act (HIPAA) have not kept pace with the important contributions AI makes to society, which may cut against traditional notions of privacy, such as data minimization. *See* Charlotte A. Tschider, *AI's Legitimate Interest*, 21 HOUS. J. HEALTH L. & POL'Y (forthcoming 2021) (describing the current state of health privacy and how it might evolve to permit greater AI data processing for the benefit of patients).

14. *See* Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1879, 1880–82 (2013) (describing the difficulty of relying on consent yet recognizing that the alternatives, such as a paternalistic approach to privacy, may not be that attractive); *see also* Neil Richards & Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. UNIV. L. REV. 1461, 1478–91 (2019) (identifying key insufficiencies in current consent models that render it ineffective); Charlotte A. Tschider, *The Consent Myth: Improving Choice for Patients of the Future*, 96 WASH. UNIV. L. REV. 1505, 1518–28 (2019) (introducing five common reasons, or myths, why consent is ineffective). It should be noted that data rights only exist in a limited number of U.S. statutes today, including the Fair Credit Reporting Act (FCRA) of 1970, 15 U.S.C. §§ 1681 *et seq*. (2020) (applicable to financial credit reporting activities); the Family Educational Rights and Privacy Act (FERPA) of 1974, 20 U.S.C. § 1232g (2020) (applicable to student records); the U.S. Privacy Act of 1974, 5 U.S.C. § 552a (2020) (applicable to government employee activities and government use of citizens' personal information); the Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of 42 U.S.C.), and its subsequent update; the Health Information Technology and Economic Health Act (HITECH) of 2009, Pub. L. No. 111-5, 123 Stat. 266 (2009) (codified as amended in scattered sections of 42 U.S.C.) (applicable to specific healthcare entities for electronic transactions involving protected health information); the Children's On-Line Privacy Protection Act (COPPA) of 1998, 15 U.S.C. §§ 6501 *et seq*. (2020) (applicable to online use of children's information for children under the age of 13); the Gramm-Leach Bliley Act (GLBA) of 1999, 15 U.S.C. §§ 6801–09 (2020) (applicable to financial institutions and their customers) and its mirror, Regulation S-P (Reg-SP), 17 C.F.R. § 248.10 (2021) (applicable to investment companies and their customers); Telephone Consumer Protection Act (TCPA) of 1991, 105 Stat. 2394 (1991) (applicable to automated dialing telephone and text marketing); and the Genetic Information Non-Discrimination Act (GINA) of 2008, Pub. L. No. 110-233, 122 Stat. 881 (2008) (applicable to genetic information collection and use); and many state privacy laws. The California Consumer Protection Act (CCPA), CAL. CIV. CODE §§ 1798.100–

International laws, including the European Union's (EU) General Data Protection Regulation (GDPR)—which has been positioned as a modern and comprehensive model for data protection—also do not effectively address AI algorithmic risks.[15] Domestic and international privacy laws generally expect, *inter alia*, that a technology creator will understand how the technology functions and be able to explain why data is being collected in some specificity, including the data elements needed to make automated decisions, *before* collecting data.[16] Such creators are also expected to determine, and should be able to explain, how an algorithm(s) makes decisions based on these data.[17] Despite the worthy goals of laws like the GDPR to prevent unlawful and unfair automated decision-making, this approach attempts to shoehorn novel technology solutions into historical privacy models.[18]

Although privacy advocates have raised issues with algorithmic decision-making for purposes of curbing discrimination and unfairness, which are serious concerns, algorithmic opacity also poses serious public safety issues and has the potential for broad and damaging social effects.[19] Legal solutions to facilitate broad algorithmic transparency seem to elude legislators, at least in part.[20] The GDPR calls for notice of automated decision-making with a corresponding "right to . . . explanation" following automated decisions, as well as the opportunity not to be subject to decisions "solely [based] on automated processing," but the United States has not followed this trajectory.[21] While it may be relatively straightforward to notify individuals of the presence of automated decision-making or profiling activities and even offer a human-based alternative decision, providing a right to explanation may be impractical, ex-

---

1798.199.100 (2018), effective in 2020, does extend some of these rights to consumers residing in California and organizations operating therein.

15. *See* Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

16. *See* Tschider, *supra* note 14, at 1526–28 (describing the temporal issues with disclosing the purposes of data use in some specificity when it is not yet understood how such data will be used in AI models prior to running algorithms, and frequently even thereafter); *see also* CHARLOTTE A. TSCHIDER, INTERNATIONAL CYBERSECURITY AND PRIVACY LAW IN PRACTICE (2017) (describing various international laws that require notice and consent).

17. GDPR, *supra* note 15, at recitals 71–72, arts. 21–22.

18. *See infra* Part II.

19. *See* sources cited *infra* note 102.

20. The opportunity to not be subject solely to automated processing is tied to whether these decisions substantially affect an individual, for example with respect to public benefits, legal consequences, or financial decisions. Of course, for many automated decisions, especially for connected devices, decisions may be embedded in the function of the device, making an alternative "human" decision an unworkable alternative. *See* GDPR, *supra* note 15, at arts. 13(2)(f), 14(2)(g), 22. Certainly, there is some debate as to the nature and extent of this requirements under the GDPR, but at least facially the GDPR does require organizations to notify individuals of an intention to profile and engage in automated decision-making.

21. GDPR, *supra* note 15, at recital 71, art. 22.

ceedingly difficult, or even impossible depending on the technology involved.[22]

The more complex the AI algorithm is, the more difficult it is to explain.[23] These complex algorithms are positioned to handle some of the most difficult human challenges, such as offering an artificial pancreas for individuals with Type I diabetes or creating dramatically safer transportation systems.[24] Ultimately, the potential benefit of employing an unexplainable algorithm may outweigh the need to directly explain how the algorithm made a decision.[25] And even where explanation is possible, it will not likely provide the kind of information needed to actually evaluate risks of unfairness, discrimination, safety, or other social impacts. As Lilian Edwards and Michael Veale have asserted: "A 'Right to an Explanation' Is Probably Not the Remedy You are Looking For."[26]

This Article seeks to identify key issues with calls for AI transparency and explainability, including the role of trade secrecy in preventing disclosure of useful information and technical complexities, and suggests alternative models for accomplishing similar goals. Specifically, this Article makes three contributions to the existing transparency and explainability literature: (1) a modern account of AI technologies, which demonstrates why previous transparency or right to explanation proposals are designed to be highly ineffective based on the technology itself; (2) an analysis of how AI algorithms establish a natural trade secret and how public policy around trade secrecy resists calls for transparency consistent with technological complexity; and (3) a recommendation of an alternative model for promoting disclosure of useful information, which balances social needs for disclosure with legal incentives to protect secrecy.

Part I introduces the modern AI-created algorithm, including the unique, complex nature of AI-enabled systems, and traverses the problem of algorithmic decision-making. In Part II, this Article discusses the theoretical goals of transparency and disclosure, including previous proposals for transparency and their relative challenges, and the unique aspects of AI systems, notably "dynamic inscrutability," which frustrates

22. *See infra* Parts I–II.

23. *See infra* Part I.

24. *What is the Pancreas? What is an Artificial Pancreas Device System?*, U.S. FOOD & DRUG ADMIN. (Aug. 30, 2018), https://www.fda.gov/medical-devices/artificial-pancreas-device-system/what-pancreas-what-artificial-pancreas-device-system; Naveen Joshi, *How AI Can Transform the Transportation Industry*, FORBES (July 26, 2019, 9:47 PM), https://www.forbes.com/sites/cognitiveworld/2019/07/26/how-ai-can-transform-the-transportation-industry/?sh=d1e593349640.

25. *See* Tschider, *supra* note 13, at 5 (advocating for interest-balancing tests to enable less restricted data processing for purposes of benefitting patients and the broader population).

26. Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 18 (2017) (evaluating the GDPR automated profiling requirements, including the right to "explanation").

many of these proposals.[27] Part III further examines the natural fit between AI algorithm-based systems and trade secrecy and identifies issues related to reliance on this system. Part III then explores the purpose of IP systems beyond the strict economic and capitalistic benefit construct to identify how a disclosure utility could also promote broad social benefits. In Part IV, this Article proposes a "deposit and disclosure" approach to sharing AI training, process, and system details that promotes third-party review and minimizes discrimination, unfairness, safety, and other potential issues in AI systems.

I. "EXPLAINING" AI IS EXCEPTIONALLY COMPLEX

AI involves algorithmic design and training, system infrastructure, and the algorithm itself.[28] The reality of modern AI is that neural networks and deep-learning applications obscure details of the AI system to such an extent that transparency is nearly impossible to achieve—even an AI's creators do not fully understand how it works.[29] AI algorithms, in their most effective state, also continue to learn and change automatically to become more effective, so even if an algorithm could be explained, it would only be explainable temporarily before the algorithm changed again.[30] Furthermore, the focus on the *algorithm* is misplaced: the methods, system, and processes that create and sustain the algorithm and its functionality actually control what the algorithm is, what is does, how reliable or fair it is, and ultimately how safe it will be.[31]

As described in Part II, the dynamic inscrutability of complex AI algorithms is buttressed with purposeful secrecy of training processes, infrastructure and AI architecture choices, data sets, and any other number of important decisions that affect the fairness and safety of AI algorithms.[32] Part I aims to describe the real complexity of these systems as designed to illustrate the broader opacity problem, which is necessary to

---

27. A second concern, "nonintuitiveness," coined by Andrew D. Selbst and Solon Barocas, and buttressed by a related concern, "specialization," is addressed in potential solutions in Parts III and IV. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–98 (2018).

28. Bob Violino, *Designing and Building Artificial Intelligence Infrastructure*, TECHTARGET (Apr. 5, 2018), https://searchenterpriseai.techtarget.com/feature/Designing-and-building-artificial-intelligence-infrastructure; Alexandre Gonfalonieri, *How to Build A Data Set for Your Machine Learning Project*, MEDIUM: TOWARDS DATA SCI. (Feb. 13, 2019), https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac.

29. Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

30. The Author introduces the turn of phrase "dynamically inscrutable algorithms" to distinguish complex AI algorithms, which are designed to self-learn and are used for higher-order decisions, from algorithms that are opaque because an organization opts not to share the details.

31. Jim Simur, *Is AI Data Driven, Algorithm Driven, or Process Driven?*, FORBES (May 6, 2019, 2:03 PM), https://www.forbes.com/sites/cognitiveworld/2019/05/06/is-ai-data-driven-algorithm-driven-or-process-driven/?sh=eaa5fd25a8d5.

32. *See infra* Part II.

understand why transparency or explanation may not actually render fair and safe AI.

A. Artificially Intelligent Systems Are Designed to Obscure, Rather Than Reveal, Decision-Making Processes

Complex AI systems are usually grouped into one of two categories: "machine learning" or "neural networking" AI.[33] Machine learning is a system that uses data to create complex algorithms via supervised and unsupervised AI learning. Supervised learning involves data scientists actively designing the function of the machine learning utility.[34] Unsupervised learning leverages a lack of prearranged structure that permits an AI utility to create its own relationships and structures between data elements.[35] For these reasons, unsupervised machine learning applications may develop higher complexity algorithms.[36] Neural networks, or deep learning systems, build on the basics of unsupervised machine learning but add additional layers of computation, further increasing computational complexity and accordingly, the algorithm itself.[37]

While some supervised machine learning algorithms may operate a lot like human-created algorithms, and potentially be explainable, they may be dynamic: they will continue to learn from new data supplied to the algorithm.[38] This means that the algorithm itself could be different in a minute or an hour after the previous decision was made. Therefore, explanation may be an impractical goal, and it may be difficult to ascertain explanation of a point-in-time decision when the algorithm has since automatically updated many times.[39]

Neural networks further increase the degree of dynamic inscrutability in their design.[40] A neural network's "neurons" add specific weightings to relationships between data elements to simulate brain function.[41]

---

33. Eda Kavlakoglu, *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?*, IBM CLOUD (May 27, 2020), https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

34. *See Supervised vs. Unsupervised Learning: Key Differences*, GURU99, https://www.guru99.com/supervised-vs-unsupervised-learning.html (last visited Feb. 23, 2021).

35. *Id.*

36. *See id.*

37. *A Basic Introduction to Neural Networks*, U. WIS. MADISON DEP'T OF COMPUT. SCIS. http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html (last visited Feb. 23, 2021); Ujjwal Karn, *A Quick Introduction to Neural Networks*, DATA SCI. BLOG (Aug. 9, 2016), https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/.

38. Although it is possible for some AI to be "locked" or self-learning while in development but then locked, or no longer self-learning for public or "production" use, this is not the typical or most accurate state for AI.

39. In contrast with code updates, AI could update every minute, every hour, every day, or every week. The benefit of AI is that it self-learns at a high rate, and it would likely not be practicable to retain all previous versions of the algorithm that results. This makes explaining the algorithm, even if it is possible, highly inefficient.

40. *See A Basic Introduction to Neural Networks*, *supra* note 37.

41. Karn, *supra* note 37.

In diagnostic medicine, certain features of a series of scanned X-ray images may play differing predictive roles for diagnosing a given condition.[42] For example, while a machine learning application might create a self-updating algorithm that directs a specific, predetermined behavior, neural networks use a group of algorithms and more complex weighting schemes to respond to any number of behaviors with highly adaptable and efficient precision.[43] For this reason, deep learning applications facilitate more complex learning applications, as might be required by artificial general intelligence or higher order robotics applications.[44]

Deep learning adopts the neural network approach to algorithmic decision-making, adding hundreds or thousands of hidden layers that simultaneously produce more accurate outputs while ultimately including increasingly obscure logic in the resulting algorithm.[45] Neurons (represented by round nodes) arranged in a deep learning schema are layered with other neurons in many-to-many orientations. (See Figure 1).[46] The function of multiple decision-making layers is that the decisions become more refined and accurate after each decision.[47] The initial layer is designed as the "input" layer and the final layer is the "output" layer.[48] Layers in between are hidden and intentionally obscured; the more layers and nodes, the more complex the algorithm becomes.[49]

---

42. Sohail Zahid, *AI in Radiology: Chest X-Rays*, SOHAIL ZAHID BLOG (Feb. 3, 2019), https://sohailzahid.com/2019/02/03/ai-in-radiology-chest-x-rays/.

43. Deep learning environments mimic deep neural relationships in the brain by using deeper, hidden layers to learn and respond to more complex tasks. These hidden and layered neurons are called multi-layer perceptrons (MLPs). MLPs operate in relation to each other and the unique circumstances posed in a given circumstance. The relative weights and behavior of MLPs are given a range, which is then validated during neural "training," and repeatedly as new decisions are made in actual, real-world use. *See What Is the Difference between Deep Learning and Usual Machine Learning*, GITHUB (Oct. 14, 2017), https://github.com/rasbt/python-machine-learning-book/blob/master/faq/difference-deep-and-normal-learning.md; HARRY A. PIERSON & MICHAEL S. GASHLER, DEEP LEARNING IN ROBOTICS: A REVIEW OF RECENT RESEARCH 3 (2017).

44. *See* PIERSON & GASHLER, *supra* note 43, at 3–4.

45. Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y 1, 5–9 (Jan-June 2016), https://doi.org/10.1177/2053951715622512.

46. *See* Karn, *supra* note 37.

47. IAN GOODFELLOW, YOSHUA BENGIO, & AARON COURVILLE, DEEP LEARNING 165 (MIT Press: 2016), http://www.deeplearningbook.org/.

48. *Id.*

49. *See Hidden Layer*, TECHNOPEDIA (Sept. 5, 2018), https://www.techopedia.com/definition/33264/hidden-layer-neural-networks.

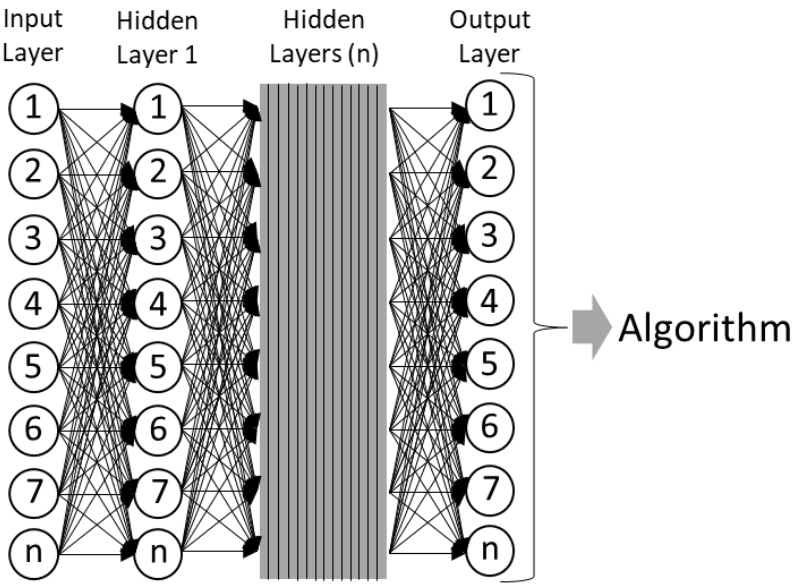

FIGURE 1. *Deep Learning Schema*

The number of layers corresponds to the complexity of the decision; more complex or refined decisions often require more layers.[50] For example, the number of layers for a high complexity algorithm, or set of algorithms, can potentially exceed 1,000.[51] These layers, the relationships between nodes within a layer, how the input and output layers are designed, and the data available to each layer are all designed by a decision or data scientist, as are many other key determinations that affect the overall fairness, reliability, and safety of such algorithms.[52]

B. Humans Make Many Unique and Creative Choices to Create Safe and Fair AI

Despite the automated and self-learning capacity of machine learning and deep learning, the effectiveness of such systems, and by extension, their relative performance, depends on human expertise and innova-

50. *See* GOODFELLOW ET AL., *supra* note 47, at 164–65.

51. *See* Muhammad Imran Razzak, Saeeda Naz, & Ahmad Zaib, *Deep Learning for Medical Image Processing: Overview, Challenges, and Future*, *in* 26 CLASSIFICATION IN BIOAPPS: LECTURE NOTES IN COMPUTATIONAL VISION AND BIOMECHANICS 323, 326–27 (Nilanjan Dey, Amira S. Ashour, & Surekha Borra eds., 2018).

52. *See, e.g.*, Harshajit Sarmah, *Google has a Chief Decision Scientist, Who is She*, ANALYTICS INDIA MAG. (Oct. 30, 2019), https://analyticsindiamag.com/data-science-google-decision-scientist/ ("Even though decision science is extensively similar to data science and even involves things like analytics, algorithms, machine learning and AI, the job of a decision scientist is more crucial. A data scientist is solely involved in extracting meaningful insights, but a decision scientist is more than that."); *Data Science for Decision Making Courses & Curriculum 2020–2021*, MAASTRICHT UNIV., https://www.maastrichtuniversity.nl/education/partner-program-master/data-science-decision-making/courses-curriculum (last visited Apr. 30, 2021) (outlining the data science instruction curriculum).

tion.[53] Machine learning best practices involve good engineering choices in code, structure of data and relevant training data sets, rule design and ongoing updates, and throughput or user and system validation.[54] Neural network choices build on machine learning choices to produce a final algorithm and decision.[55] These choices include: the orientation of neurons, depth and selection of hidden layers, the degree of vector training or self-learning, and the algorithms used to compute within these layers.[56]

An AI system is more than just an algorithm. There are at least three discrete components developed by humans that have a bearing on the fairness and safety of any AI system: (1) the method used to create a particular kind of AI, including data volume, data type, data structure, hard limits or parameters, training models, and validation or feedback loops (See Figure 2); (2) the algorithm or algorithms produced by the method that actually recommends or makes decisions; and (3) the underlying physical technology and architecture used to support the AI system, such as applications, apps, software, servers, databases, communication technologies, data transfer mechanisms, and network choices, all of which can affect the reliability and security of the entire system (See Figure 3).[57]

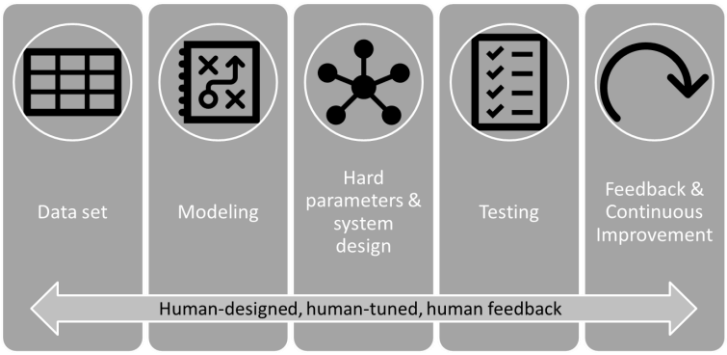

FIGURE 2. *AI Method & Design*

53. Danica Xiao, *AI vs. Rare Disease: How Human Ingenuity and Data Science Can Improve Disease Detection*, IQVIA (Feb. 28, 2020), https://www.iqvia.com/blogs/2020/02/ai-vs-rare-disease-how-human-ingenuity-and-data-science-can-improve-disease-detection.

54. *Rules of Machine Learning: Best Practices for ML Engineering*, MARTIN ZINKEVICH, http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf (last visited Feb. 24, 2021).

55. *Learning Vector Quantization (LVQ) Neural Networks*, MATHWORKS, https://www.mathworks.com/help/deeplearning/ug/learning-vector-quantization-lvq-neural-networks-1.html (last visited Feb. 24, 2021); *see also* GOODFELLOW ET AL., *supra* note 47, at 165.

56. *Learning Vector Quantization (LVQ) Neural Networks*, *supra* note 55; *see also* GOODFELLOW ET AL., *supra* note 47, at 165.

57. *See generally* Charlotte A. Tschider, *Deus ex Machina: Regulating Cybersecurity and Artificial Intelligence for Patients of the Future*, 5 SAVANNAH L. REV. 177 (2018) (describing the various aspects of AI subject to potential reliability and security concerns).

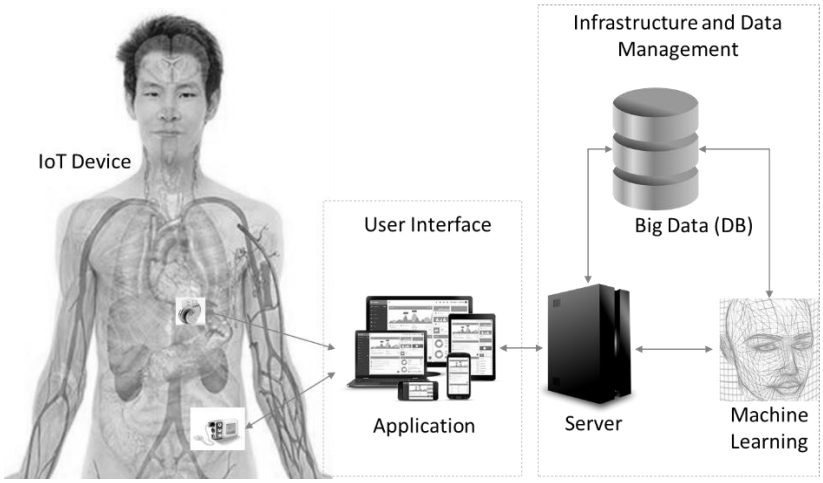

FIGURE 3. *AI Infrastructure*

1. Data Selection Affects Reliability and System Safety

Data volume and quality significantly affect an AI system's reliability and safety.[58] Of course, the volume and variety of data needed to make a system reliable and avoid these issues is highly specific to a given system's chief goals.[59] Similar to how a human mind makes assumptions without enough data inputs that may result in cognitive biases,[60] algorithms without sufficient data volume or quality data are more likely to make algorithmic assumptions, which may produce discriminatory or dangerous results.[61] For example, a home management system trained only on data from homes in warmer climates might malfunction in colder climates where fire risks are often higher in the winter. Similarly, a lifestyle-recommendation engine trained on healthy individuals could cause significant health risks for individuals with certain preexisting conditions.

Even training data—used to create and refine the initial algorithm—may include data representative of discriminatory practices.[62] By training

58. *See, e.g.*, *id.* at 178, 185–92.

59. *See* Bernard Marr, *Why AI Would Be Nothing Without Big Data*, FORBES (Jun. 9, 2017), https://www.forbes.com/sites/bernardmarr/2017/06/09/why-ai-would-be-nothing-without-big-data/#1d7677c54f6d; Joshua New, *AI Needs Better Data, Not Just More Data*, CTR. FOR DATA INNOVATION (Mar. 20, 2019), https://www.datainnovation.org/2019/03/ai-needs-better-data-not-just-more-data/.

60. *See* Jessica Stillman, *6 Cognitive Biases That Are Messing Up Your Decision Making*, INC. (Nov. 22, 2016), https://www.inc.com/jessica-stillman/6-cognitive-biases-that-are-messing-up-your-decision-making.html.

61. Matthew Stewart, *The Limitations of Machine Learning*, MEDIUM (Jul. 29, 2019), https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6.

62. SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 1–2 (2018) (describing how individual biases can be built into algorithms); *see also* Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024–25 (2018) (describing a need for validation not of the algorithm, but of the inputs and outputs). "[T]he problem is not the

AI on such data, the AI itself may codify discrimination and perpetuate it through later decisions under the guise of technical objectivity.[63] Even when such data sets do not explicitly include sensitive data elements, individuals may still experience discrimination by proxy.[64] For example, location information about an individual who frequents certain neighborhoods could potentially create an inference of religious beliefs, race, or income. If automated decisions, such as which restaurants to recommend to a given person, result in disproportionate treatment of individuals who belong to a particular group (even if the decisions themselves are seemingly innocuous), the impact itself may be discriminatory or unfair.[65]

These concepts have been discussed for some time and are certainly not new revelations. However, when data and training processes are rendered opaque through trade secrecy status, there is no meaningful opportunity to validate that such algorithms are trained to avoid discriminatory impact.[66] Fair, safe AI systems must have data sets that are large and representative enough to complete an AI system's goals, and any training processes must also be directed to the goals encoded in the system.[67]

2. AI Systems Achieve Safety and Efficacy Through Training and Feedback

For AI systems, two of the most crucial choices an AI designer must make are the training data set and the mechanisms for immediate feedback and correction.[68] For example, as Professor W. Nicholson Price has

---

black box, which is often more neutral than the human decisionmaker it replaces, but the real world on which it operates." *Id.* at 1025. Chander goes on to describe discrimination as subconscious or unconscious, which is "less likely to be encoded into automated algorithms than the human decisionmakers that the algorithms replace." *Id.* at 1028.

63. NOBLE, *supra* note 62, at 2; Paul Teich, *Artificial Intelligence Can Reinforce Bias, Cloud Giants Announce Tools for AI Fairness*, FORBES (Sept. 24, 2018, 6:00 AM), https://www.forbes.com/sites/paulteich/2018/09/24/artificial-intelligence-can-reinforce-bias-cloud-giants-announce-tools-for-ai-fairness/#bd6835e9d21f.

64. *See generally* Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1257 (2019) (describing the likelihood of discrimination when big data and highly powerful AI systems can result in discriminatory application of decisions to protected groups); Charlotte A. Tschider, *Regulating the Internet of Things: Discrimination, Privacy, and Cybersecurity in the Artificial Intelligence Age*, 96 DENV. L. REV. 87, 98–100 (2018) (describing the likelihood of discrimination in big data collection and analyses and the potential for disparate treatment and discriminatory impact through the increased use of algorithms).

65. *See* Tschider, *supra* note 64, at 101–02.

66. *See* NOBLE, *supra* note 62, at 1–3. Noble dubs this "algorithmic oppression," wherein an individual cannot know what goes into an algorithm yet may simultaneously be oppressed by it. Ultimately, the goal of most due process recommendations is to provide redress for oppression by offering the ability to contest a result and receive an explanation of the algorithm's decision. As this Article will describe and Noble intimates, the problem may not be the algorithm at all, but rather the logic and data that created it, which are selected and designed by humans. *Id.* at 29.

67. Mark A. Lemley & Bryan Casey, *Fair Learning*, TEX. L. REV. 101, 120–21 (forthcoming 2021).

68. *See* Daryna Kacherovska, *How to Prepare Training Data for Better AI?*, TOWARDS DATA SCI. (Jul. 30, 2019), https://towardsdatascience.com/how-to-prepare-training-data-for-better-ai-43c52e782c8a.

identified regarding health diagnostics, if a system trains on data from hospitals with a high degree of resources—such as the newest technologies and the most highly trained practitioners—the model the AI system creates will be oriented towards high-resource use and may not be as effective as one trained on low-resource environments.[69] To avoid this disparity, training data should be representative of the population or community where the AI might be used.[70]

In another example, AI facial recognition systems may have only been trained on light skin but the live sensors and corresponding AI cannot recognize the presence of darker skin tones.[71] When AI designs cannot even detect the presence of an individual due to the color of their skin, it is not difficult to imagine how AI design, in terms of training data sets, could affect AI system fairness or safety.[72] When such systems do not involve effective feedback loops that promote correction and ongoing training in the AI algorithm's development phase, an AI system risks entrenchment in the fairness and safety problems codified through non-representative data sets.[73]

3. Task and Goal Design Informs AI Functionality and Reveals Potential Problems

AI developers create AI applications with specific tasks in mind, and these tasks are associated with certain attendant goals and constraints.[74] For example, a maps application might be given the task "arrive in Seattle as soon as possible," or a self-driving car might be given the task "maximize power efficiency during this trip," but both need attendant constraints, such as "avoid accidents" or "keep interior temperature constant." Although some AI may be designed as extensible and

---

69. W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J.L. & TECH. 66, 66–68 (2019).

70. *Id.* at 92.

71. Tom Simonite, *The Best Algorithms Struggle to Recognize Black Faces Equally*, WIRED (July 22, 2019), https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/.

72. Consider, for example, a self-driving car that does not recognize pedestrians who have a certain skin pigment. These cars could pose alarming safety hazards.

73. Ian Xiao, *The Last-Mile Problem of AI*, TOWARDS DATA SCI. (Sept. 9, 2018), https://towardsdatascience.com/fixing-the-last-mile-problems-of-deploying-ai-systems-in-the-real-world-4f1aab0ea10; *see also* Price II, *supra* note 69, at 110. Although for some AI, such as those used by experts, human expert validation may prevent some issues; the entire point of AI is to not need human experts to validate. When systems depend on human expert validation, but experts are not available, the system may still be vulnerable to existing issues. Further, over time, trained incapacity may result in medical professionals relying on AI to the extent that external validation is not possible. For these reasons, it is centrally important that feedback loops are used during the training process and systems are evaluated by experts for an extended period after the AI is released to the public.

74. *See* Kristinn R. Thórisson, Jordi Bieger, Thröstur Thorarensen, Jóna S. Sigurðardóttir, & Bas R. Steunebrink, *Why Artificial Intelligence Needs a Task Theory – and What it Might Look Like*, ARTIFICIAL GEN. INTEL. 4–5 (2016), https://www.researchgate.net/publication/304459314_Why_Artificial_Intelligence_Needs_a_Task_Theory_-_And_What_It_Might_Look_Like.

applicable to a variety of different circumstances, they must be fit, adjusted, tested, and able to receive feedback to adequately fulfill the key tasks posed.[75] Furthermore, the technical infrastructure selected must support these tasks and their required capacity (e.g., data volume, processing speed, redundancy, or other performance criteria).[76]

Ultimately, recipe-like requirements cannot eliminate potential risks because an AI system that is safe for one application could be completely unsafe for another.[77] One AI system that does not produce discriminatory results in one context may produce discriminatory results in another.[78] An AI system that supports medical diagnosis will be designed differently, both in system and in process, than an AI system for self-driving cars, which has very different tasks to perform.[79] An AI system designed to read mammogram X-ray images may not work as effectively on images from an ultrasound machine. Industry best practices, such as those developed by the National Institute of Standards and Technology or the International Standards Organization, can reduce potential risks.[80] However, these standards cannot prevent potential discrimination or safety issues, as these standards cannot possibly anticipate every task, goal, and constraint specific to a given system.[81] As such, AI is inherently fit for purpose.

---

75. B.J. Copeland, *Artificial Intelligence*, ENCYCLOPAEDIA BRITANNICA (Apr. 11, 2019), https://www.britannica.com/technology/artificial-intelligence.

76. *Id.*

77. This is simply the nature of computerized systems, and this argument has been made frequently in the cybersecurity area as well. *See* David Thaw, *The Efficacy of Cybersecurity Regulation*, 30 GA. ST. U. L. REV. 287, 299 (2014) (describing the role of "management-based regulation" wherein cybersecurity is not advanced through direct requirements but rather through "conditions that must be met during the planning stage of a productive process—before manufacture of a product or provision of a service begins."). Management-based regulation embraces the fluid nature of technology, acknowledging that dynamic systems might have different answers to the same challenge, even though certain principles might be similar from application to application. *Id.*

78. *Id.*

79. Daniel Faggella, *Machine Learning for Medical Diagnostics—4 Current Applications*, EMERJ (Mar. 14, 2020), https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/; Alex Davies, *The WIRED Guide to Self-driving Cars*, WIRED (Dec. 13, 2018), https://www.wired.com/story/guide-self-driving-cars/; Tonya Riley, *Get Ready, This Year Your Next Job Interview May Be with an A.I. Robot*, CNBC (Mar. 13, 2018), https://www.cnbc.com/2018/03/13/ai-job-recruiting-tools-offered-by-hirevue-mya-other-start-ups.html. AI utilities are designed specifically to fulfill or maximize certain codified goals. For example, medical diagnostics might be designed to identify breast cancer from mammogram images, while a self-driving car might be designed to identify and follow traffic signals, and an AI employment application might be designed to detect false narratives or measure reliability. *See* Jim Guszcza, *Smarter Together: Why Artificial Intelligence Needs Human-centered Design*, DELOITTE INSIGHTS (Jan. 22, 2018), https://www2.deloitte.com/insights/us/en/deloitte-review/issue-22/artificial-intelligence-human-centric-design.html.

80. NAT'L INST. OF STANDARDS & TECH., *AI Standards*, NIST (June 9, 2020), https://www.nist.gov/topics/artificial-intelligence/ai-standards; Robert Bartram, *The New Frontier for Artificial Intelligence*, INT'L STANDARDS ORG. (Oct. 18, 2018), https://www.iso.org/news/ref2336.html.

81. The "privacy by design" or "privacy by default" movement has addressed this issue from a privacy perspective. It is enshrined in the GDPR but was first proposed and described by Privacy Commissioner Ann Cavoukian for Canada's Ontario province in the late 1990s. *See* GDPR, *supra*

Technology is fluid, and technical standards, when established by a standards bearer, are not.[82] For AI, standards alone cannot solve potential unfairness and safety issues. Rather, as many astute scholars including Danielle Keats Citron, Frank Pasquale, Solon Barocas, Andrew Selbst, Danah Boyd, Kate Crawford, Jason Schultz, and Rashida Richardson have contended, transparency in the form of explanation or responsive due process review may reveal unfair practices and the threat of review may promote better design practices.[83]

4. Physical System Design and Architecture May Introduce Substantial Risk to Individuals

When AI utilities are executed over remote systems, as is usually the case with Internet of Things (IoT) devices, technical architects must design how devices transmit data to back-end infrastructures depending on how the AI system will be used by these devices.[84] Although AI may be designed to be extensible, the structure and system must be adjusted when used for a specific application: the process must be customized when using a commercial AI platform.[85] This is usually called "AI as a Service," and reduces investment in creating specific fit for purpose AI.[86] For example, medical AI that diagnoses breast cancer will likely involve connectivity to a back-end infrastructure hosted by a third party. Without consideration of cybersecurity, AI systems could produce discriminatory or inaccurate results.[87]

---

note 15, at art. 25; Ann Cavoukian, *Privacy by Design*, INFO. & PRIV. COMM'R OF ONT. (Jan. 2011), https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

82. *See* Thaw, *supra* note 77, at 299.

83. *See* Selbst & Barocas, *supra* note 27; *see also* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 99, 127 (2014) (describing how predictive analytics can actually create new privacy harms, and how a responsive process to redress these harms is important); Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, *in* 2019 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAT*) 59 (2019) (describing the challenges in evaluating "fairness" at various points in an AI system's development); Rashida Richardson, Jason M. Schultz, & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 192, 218 (2019) (describing how using existing police data for automated systems provides a confirmatory feedback loop that is nonetheless discriminatory because it is based on discriminatory data). *See* sources cited *infra* note 102.

84. A plethora of choices, including the system, process to design the AI, and overall use of the AI (i.e., device, diagnostics, or automated decisioning), must be tailored specifically to the algorithm's use. *See* Violino, *supra* note 28.

85. Margaret Rouse, *Artificial Intelligence as a Service (AIaaS)*, TECHTARGET (Dec. 2017), https://searchenterpriseai.techtarget.com/definition/Artificial-Intelligence-as-a-Service-AIaaS.

86. Joe McKendrick, *Get Ready for AI as a Service*, FORBES (Mar. 16, 2020), https://www.forbes.com/sites/joemckendrick/2020/03/16/get-ready-for-ai-as-a-service/?sh=315f48f7280b; *see also* Lemley & Casey, *supra* note 67, at n.112 (citing Jessica L. Gillotte, *Copyright Infringement in AI-Generated Artworks*, 53 U.C. DAVIS L. REV. 2655, 2684 (2020) (arguing "that AI training data sets are 'highly transformative' because the purpose of use is different"; the concept of purpose of use illustrates the changing dynamics when one model is used in another scenario).

87. Consider a scenario where an organization uses a common platform but fails to test it or assess its security status. It could mean that either the system produces discriminatory results or the

Because many AI systems are designed by start-ups and later licensed or sold to larger companies, AI may not be designed and optimized for broad use across populations and geographies.[88] These systems may be vulnerable to security threats simply because AI design in a lab did not account for broad-scale attacks.[89] The most dangerous aspect of poor design is that—because AI systems are not easily explainable—it is more likely that unauthorized data changes will cause major issues.[90] For example, a home thermostat could be updated or given direction by an AI algorithm in a "smart" home. If the data is changed in an unauthorized manner, the system giving instructions to the home thermostat based on that data could increase the limits of a thermostat beyond its typical threshold, causing a furnace to catch fire.[91] Without understanding how the system is designed or whether a security system has been implemented, it is impossible to know when serious issues arise.[92]

Of course, the fit for purpose aspect of AI combined with the inherent complexity and changeability of its algorithms poses unique challenges for *ex ante* oversight; what is a perfectly reasonable design for one AI system may not appropriately acknowledge issues for another.[93] Privacy and discrimination scholars, as well as product safety advocates, have proposed a move toward AI algorithmic explainability for purposes of determining and preventing unanticipated negative consequences for natural persons using or affected by these technologies.[94] However, both the inherent inability to make dynamically inscrutable algorithms continuously explainable and the desire for organizations to protect their creations will likely frustrate the notion of legally mandated transparency.[95]

---

system's poor security results in privacy or physical injury issues. *See* Tschider, *supra* note 57, at 179, 192–96, 198–201 (describing issues in AI regulation by the U.S. Food and Drug Administration (FDA) specifically related to cybersecurity cyber-kinetic attacks of medical devices).

88. *The Race for AI: Here Are the Tech Giants Rushing to Snap Up Artificial Intelligence Startups*, CB INSIGHTS (Sept. 17, 2019), https://www.cbinsights.com/research/top-acquirers-ai-startups-ma-timeline/; Gabriela Barkho, *Funding for Artificial Intelligence Startups Reaches Record High in 2019*, OBSERVER (July 26, 2019), https://observer.com/2019/07/artificial-intelligence-startups-funding-record-high-2019/.

89. *See* Tschider, *supra* note 57, at 189–91, 208–09.

90. *See* Tschider, *supra* note 64, at 118–20.

91. Smart-home applications are already being targeted by cybercriminals. *See* Anthony Spadafora, *Smart Home Devices Are Being Hit with More Cyberattacks Than Ever*, TECHRADAR.PRO (Oct. 15, 2019), https://www.techradar.com/sg/news/smart-home-devices-are-being-hit-with-more-cyberattacks-than-ever. AI systems only make the visibility of unauthorized changes that much more difficult to detect. *See* Tschider, *supra* note 57, at 187–88.

92. Ben Buchanan, *The Future of AI and Cybersecurity*, THE CIPHER BRIEF (Oct. 30, 2019), https://www.thecipherbrief.com/column_article/the-future-of-ai-and-cybersecurity.

93. For example, many training data sets might contain information that is more or less useful for specific uses and may result in less useful functionality. *See* Drew Roselli, Jeanna Matthews, & Nisha Talagala, *Managing Bias in AI*, *in* COMPANION PROCEEDINGS OF THE 2019 WORLD WIDE WEB CONFERENCE (2019), https://people.clarkson.edu/~jmatthew/publications/ManagingBiasInAICAMERAREADY.pdf.

94. *See infra* Part II.

95. *See infra* Part III.

II. EXPLAINABILITY MAY NOT ACHIEVE IMPORTANT SOCIAL GOALS

The "transparency" concept in relation to black box decision-making was initially introduced to respond to opaque algorithmic decision-making.[96] For the past fifteen years, scholars have positioned transparency, or explanation, as the solution to a variety of potential issues—from unfairness or discrimination to due process concerns.[97] The algorithms of today, however, are both opaque by organizational choice and by technical reality: organizations may choose to keep certain information about systems secret, such as the infrastructure of AI systems and methods to create AI algorithms, and the algorithm itself may be technically inscrutable, a natural trade secret.[98] The demand for IP protection and technical limitations may make previous calls for complete transparency undesirable and, occasionally, nearly impossible.[99]

A. The Algorithmic Black Box, Whether AI or Human-Created, Will Likely Impact Fairness and Safety

Algorithms overall, whether human-designed or AI-produced, are described as a black box problem, meaning that the algorithm itself is opaque, rather than transparent, whether the result of dynamic inscrutability in the AI algorithm or a purposeful non-disclosure of a trade secret or confidential business information.[100] Black box problems exist when, generally construed, opacity has a significant impact on natural persons or consumers.[101] Legal scholars including Danielle Citron, Frank Pasquale, Tal Zarsky, Kate Crawford, Jason Schultz, Andrew Selbst, Nicholson Price, Roger Allan Ford, and myself have discussed these potential black box algorithm issues, including issues related to privacy, fairness, and safety.[102] Collectively, these scholars have identified the risks associated

---

96. *See* sources cited *infra* note 102.

97. *See* sources cited *infra* note 102.

98. *See* Selbst & Barocas, *supra* note 27, at 1091–93; *see generally* Jeanne C. Fromer, *Symposium*: *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. Rev. 706, 719–20, 722–24 (2019) (describing the role of trade secrecy in AI).

99. *See infra* Part III.

100. The concept of a "black box" has long existed prior to discussions about algorithms, but scholars have positioned the black box as specific to algorithms that are not explainable or transparent to the individuals they affect. *See infra* note 102; Adrian Swinscoe, *Customer Experience, Opaque AI and the Risk of Unintended Consequences*, FORBES (Jul. 27, 2017), https://www.forbes.com/sites/adrianswinscoe/2017/07/27/customer-experience-opaque-ai-and-the-risk-of-unintended-consequences/?sh=7e3ac06e6963.

101. Swinscoe, *supra* note 100.

102. *See, e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1253–55, 1296–97, 1308–13 (2007) [hereinafter *Technological Due Process*] (identifying government-induced and judicial-technological due process concerns, establishing the role of computer engineers as rule-making authorities without appropriate notice and comment periods, and proposing rules and standards for addressing these concerns, including explanation of the extent a decision relies on automated decision-making, the actual code used for decision-making, public review and involvement in system design, and rigorous testing); Frank A. Pasquale, *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235, 236–40, 244–50 (2011) [hereinafter *Restoring Transparency*] (noting the primacy of trade secrecy for commercial business as an alterna-

with automated decision-making, whether AI-enabled or not.[103] Without additional interventions, AI systems will likely result in unfair automated determinations, potentially resulting in discriminatory impact, and AI systems will likely not be as safe.[104]

For many scholars, transparency—with disclosure as one of its mechanisms—is positioned as part of the cure for any number of algorithmic ills, especially with a standard-bearing or reliable organization's imprimatur.[105] Core to the concept of transparency is the opportunity for

---

tive for patent disclosure, describing the specific circumstances where transparency may outweigh secrecy, and advocating for greater transparency; noting the potential for gaming the system when transparency increases); Crawford & Schultz, *supra* note 83, at 109, 119–20, 122–23, 125–28 (identifying the potential for due process implications in big data collection of personal information and proposing technological due process considerations for big data decisions, such as providing notice of data use from external sources and resolution via a "data arbiter" through the Federal Trade Commission); Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 18–26 (2014) [hereinafter *Scored Society*] (proposing a procedural due process framework for algorithmic scoring and decisions in credit decisions involving algorithmic disclosure to the Federal Trade Commission and licensing agencies); Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375, 1378–81, 1384–96 (2014) (expanding Citron and Pasquale's concerns around due process to discrimination in business, government, and social circumstances, specifically identifying the potential for explicit and implicit discrimination and bias related to "tainted data"); Frank Pasquale & Danielle Keats Citron, *Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society*, 89 WASH. L. REV. 1413, 1417 (2014) [hereinafter *Promoting Innovation*] (acknowledging the integrated nature of protected classes within scoring and decision-making systems); Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7(4) INTL. DATA PRIV. L. 233, 6–8 (2017) (describing the debate over whether the GDPR positively requires a right of explanation); W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 429, 432–37, 440–42, 460, 465–66 (2015) [hereinafter *Black-Box Medicine*] (introducing the unique impacts of algorithmic decision-making in the healthcare sector and the need to validate predictive analytics); Roger Allan Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1, 29–37 (2016) [hereinafter *Privacy and Accountability*] (addressing privacy issues through examining collection, use, and disclosure through third-party verification techniques managed by the Food and Drug Administration); W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 465–70 (2017) [hereinafter *Regulating Black-Box Medicine*] (proposing an algorithmic disclosure model to the Food and Drug Administration and collaborative feedback from users); Frank A. Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO STATE L. REV. 1243, 1250 (2017) [hereinafter *Toward a Fourth Law of Robotics*] (examining Jack Balkin's proposed laws of robotics, including the concept of an information fiduciary, and arguing that "it's important to maintain deontological patterns of justification in the technology world to complement the utilitarianism of cost-benefit analysis"; for example, in circumstances involving discrimination, cost-benefit analysis may not be appropriate); *see also* Edwards & Veale, *supra* note 26, at 48–51, 77–79.

103. *See* sources cited *supra* note 102.

104. It should be noted that some industries do have mechanisms for *potentially* reviewing these technologies from a safety perspective. The FDA, for example, reviews device submissions, including those involving AI. However, the FDA does not expand across the entire AI industry. And, the current FDA system has not developed mandatory AI requirements, either through the Code of Federal Regulations or comprehensive Guidance. *See* Charlotte A. Tschider, *Preempting the Artificially Intelligent Machine*, BYU L. REV. 1, 19–20, 39 (forthcoming 2021).

105. *Id.* at 52–53; Pasquale, *Restoring Transparency*, *supra* note 102, at 236. The government first communicated about the transparency and objectivity of systems to quell concerns over law enforcement and IRS scoring systems. However, the rise of privatized systems has replaced the concept of "legitimation-via-transparency" with "reassurance-via-secrecy." *Id.* at 237; *see also Regulating Black-Box Medicine*, *supra* note 102, at 465 (describing the potential for third-party verification); *Privacy and Accountability*, *supra* note 102, at 35 (promoting gate-keeping mechanisms for data use and system approvals).

inspection and testing—either prior to algorithm use or as a response to claimed injury—as an extension of computational governance or within a fiduciary relationship between technology providers and those affected by the technology.[106] The challenge with AI is what Jack Balkin calls the "substitution effect," or the effect of substituting human decision-making with computing solutions.[107] In short, we expect computers to function like humans with all of the capabilities of computers, including being able to "explain itself."[108]

Under the substitution effect, computers produce both better and worse results, while also existing both within and away from traditional social constructs and rules.[109] This substitution effect, combined with the power of AI—the ability to quickly identify complex relationships among seemingly disparate data elements—complicates its regulation.[110] AI is *not* human, yet it may affect human relationships and prospects.[111] And the very engine that complicates its regulation is the same mechanism that will likely improve process efficiency, efficacy, reliability, safety, and capacity, and reach into various disciplines: from healthcare to transportation, environmental use to farming.[112] However, the dangers associated with poor design or cyberattack are myriad and alarming.[113]

On a broad-scale basis, the GDPR first codified the concept of automated system transparency via organizational notice and review requirements for organizations doing business with EU data subjects or natural persons.[114] Other countries have begun to replicate the GDPR:

---

106. Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO STATE L.J. 1217, 1226 (2017).

107. *Id.* at 1224.

108. Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES MAG. (Nov. 21, 2017), https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

109. *Id.*

110. Indeed, the power of AI and its possibilities in any number of industries has caused a rapid expansion of the technology while simultaneously leaving us no closer to clear boundaries and clear rules. Its use across sectors, for example, means that direct regulation of AI would likely mean some federal omnibus law, which so far has not been too appetizing for Congress. *See* Quora, *Should Artificial Intelligence Be Regulated?*, FORBES (Aug. 31, 2017, 2:15 PM), https://www.forbes.com/sites/quora/2017/08/31/should-artificial-intelligence-be-regulated/?sh=1142400331d7 (describing the challenges in regulating AI).

111. *See Technological Due Process*, *supra* note 102, at 1281; *Scored Society*, *supra* note 102, at 2–6; Crawford & Schultz, *supra* note 83, at 97–100; *Promoting Innovation*, *supra* note 102, at 1422–24; Zarsky, *supra* note 102, at 1405–12 (describing the effects on individuals in a "scored" society).

112. Magnimind, *Invaluable Societal Benefits of AI*, MEDIUM (Mar. 19, 2019), https://becominghuman.ai/invaluable-societal-benefits-of-ai-2ed62f7a653f (describing potential societal benefits while also noting that "transparency is the key"). These benefits, however, do not necessarily account for labor distribution. For a piece on labor distribution, see Julia M. Puaschunder, *Revising Growth Theory in the Artificial Age: Putty and Clay Labor*, 8 ARCHIVES BUS. RSCH. 65, 77–78 (2020) (describing labor dynamics in the AI age).

113. *See* Tschider, *supra* note 64, at 116–21 (describing cyberkinetic attacks and privacy-compromising attacks).

114. *See* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 189–92 (2019).

first by Brazil, the largest economic area in South America.[115] The GDPR hinges individual civil rights on bookended requirements: notice that automated decision-making is occurring and the ability for data subjects to enforce their rights to explanation and of objection to automated decisions about them.[116] These rights to explanation presumably are "sufficiently interpretable [by] . . . a human arbiter."[117] The GDPR is framed in reference to the civil liberties of natural persons and their corresponding information.[118] From this perspective, the GDPR focuses on unfairness of automated decision-making that may impact the rights and freedoms of natural persons.

1. Transparency for Safety

However, the concerns around AI are more extensive than the role automated decision-making plays in relation to personal information. Safety issues, for example, may have nothing to do with personal information but can still affect consumer health and property.[119] Products incorporating AI include electronics both in traditional electronic devices and other digital IoT, as well as support process for product development, supply chain, and customer insights.[120] Product safety, however, is largely left to responsive measures after issues arise, and many of these issues are unforeseeable from the perspective of the consumer.[121] Globally, regional governmental entities like the EU are considering the best method for regulating AI products.[122]

---

115. Lei Geral de Proteção de Dados Pessoais [The General Data Protection Law], Law No. 13,709/2018. In the financial, credit, and insurance sectors, laws have required greater transparency in decisions rendered, some of which may be automated. However, the GDPR is the first to require an explicit, up-front notice of automated decisioning and profiling, combined with a right to understand the results of such decisions and object to such profiling and subsequent decisioning by automated means (when such decisions substantially impact other individual rights). *See* GDPR, *supra* note 15, at recital 71, arts. 13–14, 22.

116. GDPR, *supra* note 15, at recital 71, arts. 13–14, 22.

117. Emre Bayamlıoğlu, Transparency of Automated Decision in the GDPR: An Attempt for Systemisation 39 (Nov. 29, 2018) (unpublished manuscript) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3097653).

118. *See* Kaminski, *supra* note 114, at 204 n.83.

119. *See, e.g.*, Tschider, *supra* note 64, at 109 (describing that cybersecurity safety issues in the IoT are not necessarily privacy issues).

120. *See* IBM INSTIT. FOR BUS. VALUE, THE COMING AI REVOLUTION IN RETAIL AND CONSUMER PRODUCTS (2019); Karan Bedi, *How Artificial Intelligence is Reinventing Consumer Electronics Segment*, ENTREPRENEUR (Feb. 21, 2019), https://www.entrepreneur.com/article/328400; Sofiane Boukhalfa, *IoT and AI Combine to Dominate Consumer Electronics*, PRESCOUTER (May 2017), https://www.prescouter.com/2017/05/iot-ai-consumer-electronics/; Ron Schmelzer, *Making the Internet of Things (IoT) More Intelligent with AI*, FORBES (Oct. 1, 2019, 8:15 AM), https://www.forbes.com/sites/cognitiveworld/2019/10/01/making-the-internet-of-things-iot-more-intelligent-with-ai/#32b53bdbfd9b.

121. Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1329–32 (2020).

122. Gabriele Mazzini, *A System for Governance for Artificial Intelligence through the Lens of Emerging Intersections between AI and EU Law*, *in* DIGITAL REVOLUTION – NEW CHALLENGES FOR LAW 245, 252–57 (Alberto De Franceschi & Reiner Schulze eds., 2019) (describing the potential for product safety legal questions as an extension of risk-balancing efforts for AI: AI may be worth

2. Transparency for Fairness

A key issue in algorithmic decision-making is the degree to which such decisions may unfairly impact individuals with certain characteristics or with respect to certain groups. Fairness is not only defined by traditionally identified categories of discrimination but by fairness to a certain individual and their preferences or lifestyle.[123] U.S. law has not specifically addressed the latter definition of fairness, focusing instead on more traditional definitions of discrimination based on categories like race, religion, gender, or disability.[124] Under historical laws that incorporate these categories, typically legal restrictions are tied to intentional or reckless decisions based on these categories; most regulation has equated liability with intent.[125] The California Consumer Protection Act (CCPA), modeled in part after the GDPR, has prohibited discrimination in goods and services (including pricing and availability) when consumers have exerted their privacy rights.[126] Despite this unusual prohibition, even the CCPA has not prescribed fairness in algorithmic decision-making.[127] However, the duties identified as part of the "information fiduciary" role, which is gathering recent attention, may offer an opening for other incentive-based solutions even when legal mandates fail.[128]

B. Transparency Has Been Positioned as the Means for Reinforcing Responsibility

Although mandating algorithmic transparency may appear to solve these black box problems, it may be an inefficient and ineffective means to accomplish the desired outcome. As an initial matter, a blanket call for algorithmic explainability oversimplifies algorithmic complexity and dynamic inscrutability.[129] Moreover, it presumes that what is "revealed" upon explanation would actually be useful and *intuitive*.[130] Calls for explainability fail to appreciate a crucial aspect of AI systems: when it comes to fairness and safety, more than just the algorithm matters. Distinctly human choices, such as data set size, source, and variety, training

---

investment for the potential benefits, and the law should balance these against the potential risks inherent in these technologies).

123. *See* Prince & Schwarcz, *supra* note 64, at 1276–82 (describing how indirectly identifiable data can inform proxies for potential discriminatory categories); *see also* Selbst et al., *supra* note 83, at 59–63 (describing common pitfalls in fairness engineering for machine learning).

124. *See* Prince & Schwarcz, *supra* note 64, at 1276–82.

125. *See Toward a Fourth Law of Robotics*, *supra* note 102, at 1247.

126. California Consumer Protection Act of 2018, CAL. CIV. CODE § 1798.125 (West 2020).

127. *Id.*

128. Although the Author does not necessarily support legal recognition of an information fiduciary, the concept is not incompatible with potential solutions included here. *See infra* Part III.

129. *See supra* Part I.

130. *See supra* Part I; Selbst & Barocas, *supra* note 27, at 1096–98; Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 4 (2017) (arguing that the push for transparency is misguided because it misunderstands the nature of the algorithms at stake); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Applications to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 981–84 (2018).

protocol and feedback loops, and infrastructure can fundamentally alter the fairness and safety of AI systems.[131]

Choices executed to create AI are essential to understanding its relative decision-making, including choices both in designing the AI algorithm and architecting its overall system and structure. For example, in designing machine learning applications to avoid fairness issues, Andrew Selbst et al., have identified five traps relating to how AI systems are built, and at least two relate to broad technology decisions as described in Part I: (1) failure to model fairness over an entire system, and (2) failure to understand how the portability of algorithms between contexts may be misleading, inaccurate, or harmful in another context.[132]

AI also has complicated privacy-specific data considerations including large-volume data aggregation, data repurposing and recombination, data store amendment and supplementation, and opaque data processing.[133] Although privacy issues are certainly exacerbated in AI systems, dynamic inscrutability and intentional secrecy of data sets, training processes, and infrastructure choices impact much more than privacy.[134] In contrast with human-created algorithms, where opacity results from a lack of disclosure, a lack of AI algorithmic transparency results specifically from technical complexity, which even technical expertise may not overcome.[135] In machine learning, algorithms are created from dynamic data, rather than designed by data scientists, so that an algorithm might change at any given point in time.[136] Further, the complexity of a data source might mean an algorithm is lengthy and complex, to the point where even data scientists may not fully comprehend its effect.[137] In neural networks, the hidden nature of computational layers and their respective weights, as well as the compound nature of multiple simultaneously computing algorithms, makes any resulting algorithmic direction exceptionally difficult to ascertain.[138] In this way, there is a better term than *opacity* when applied specifically to AI algorithmic complexity: *dynamic inscrutability*.[139] Dynamic inscrutability complicates both preventative

131. *See supra* Part I.

132. *See* Selbst et al., *supra* note 83, at 60–61.

133. *Privacy in the Age of Data Aggregation*, MEDIUM (Mar. 7, 2018), https://medium.com/vetri/privacy-in-the-age-of-data-aggregation-15fc87328209.

134. *Id.*

135. Aaron M. Bornstein, *Is Artificial Intelligence Permanently Inscrutable?*, NAUTILUS (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable.

136. *See* Vincenzo Lomonaco, *Why Continual Learning Is the Key Towards Machine Intelligence*, MEDIUM (Oct. 4, 2017), https://medium.com/continual-ai/why-continuous-learning-is-the-key-towards-machine-intelligence-1851cb57c308.

137. *See* Knight, *supra* note 29.

138. *See* Bornstein, *supra* note 135.

139. *See* Selbst & Barocas, *supra* note 27, at 1089–91. The Author adopts this terminology to represent the technical challenges leading to opacity. Instead of algorithms that are opaque until an organization chooses to reveal them, dynamic inscrutable algorithms are a result of advanced AI techniques and are distinct from non-intuitiveness, as described by Selbst and Barocas. *Id.* In addition to what Selbst and Barocas usefully delineate in the AI space, the nature of deep learning and

and responsive approaches to AI transparency, frustrating models that previously applied to human-created algorithms.[140]

Dynamic inscrutability creates significant issues for the reliability of decisions and recommendations, and it may perpetuate safety issues. Understanding whether an AI decision is reliable is crucial to significant decisions, such as medical diagnoses, employment, or eligibility for public services. A lack of explanation in these situations makes it difficult to determine the right responsive action.[141] For example, for medical diagnoses, medical professionals might benefit from understanding how a diagnostic algorithm reaches decisions because they can then assess whether all potential factors for a particular disease or disorder were considered. In employment decisions, a candidate might be subject to discrimination when both the employer and candidate cannot evaluate how the decision was made. For public services eligibility or any other government decision, a lack of access to the details of the decision may create barriers to responsive action, such as appeal.[142]

Dynamic inscrutability may also perpetuate safety issues, both prior to product availability and for product liability cases or other torts.[143] When AI is launched, a lack of peer review or other algorithmic inspection could result in unsafe AI-enabled products being downloaded by unsuspecting consumers, causing injury or property damage.[144] After injury or property damage, these same consumers pursuing legal action may not be able to demonstrate clearly why an organization should be liable, because the algorithm may be so inscrutable it is not possible to

---

neural networks, as described, requires an ongoing dynamic learning capability to effectively solve dynamic and changing real-life problems, such as diagnosing a health condition, navigating a self-driving car, or autonomously operating an electrical grid. *Id.* In these cases, it is not as simple as requiring algorithmic explanation because an organization or data scientist cannot readily explain the algorithm resulting from hundreds or thousands of layers of independent calculations and sub-algorithms. "Inscrutability" has been used for some time to describe complex AI functionality that may be impossible to explain. "Dynamic" represents the ephemeral nature of continuously changing algorithms.

140. *See Technological Due Process*, *supra* note 102, at 1281–82; *see also* Crawford & Schultz, *supra* note 83, at 96–99. Consider, for example, how due process recommendations, as described by Danielle Citron, Kate Crawford, and Jason Schultz, rely on being able to provide *some* explanation. When such explanation is (1) difficult if not impossible to find, and (2) the logic underlying the explanation is dynamically changing, these notions of due process may be frustrated, absent something more. It should be noted that considerable effort is being made to develop technical methods to explain such algorithmic decisions. However, such explanations may nevertheless only apply in a given scenario and do not reflect decisions after a system change has taken place. Storing every decision that has ever taken place under any version of an algorithm is likely to demand substantial storage space, as well. Finally, the output of such analyses may not be tremendously useful or helpful given the complexity of calculation for higher-order AI. All of this necessitates another way to provide transparency.

141. *See Technological Due Process*, *supra* note 102, at 1281–82; *see also* Crawford & Schultz, *supra* note 83, at 87–98, 100–03.

142. *See* Crawford & Schultz, *supra* note 83, at 117–21.

143. *See* Selbst, *supra* note 121, at 1349–50.

144. *Id.* at 1330.

delineate how the injury occurred or if an imperfection in the system was the proximate cause of the injury.[145]

However, the real origin of potential discrimination or injury lies with the entire system's design, not just the resulting algorithm. Poorly designed systems will likely result in more issues, while responsibly and ethically designed systems will likely not.[146] As Daniel Susser has described, "invisible architecture" is a core concern for AI systems, not only the dynamically inscrutable algorithm.[147] In fact, the specific form of the information presented to a system user constrains what a user might expect about its attendant risks.[148] For example, a simple interface might deceive a user into believing that a simple AI-enabled system lies behind the interface. Over time, users might not even focus on changes to the technology, but instead may focus primarily on the tasks they plan to execute as the machine churns behind the scenes.[149] The risk, therefore, is that as users become more adept at using these technologies, there is actually more potential for manipulation and discriminatory or unsafe outcomes because the user becomes comfortable or perhaps even forgets that a complex computer system is actually influencing their behavior.[150]

For these reasons, transparency or disclosure in some form is core to avoiding potential unfairness and safety issues.[151] However, calls for transparency ignore critical issues for how organizations function to protect their resources. Specifically, calls for transparency ignore the importance of IP protection to businesses, especially start-ups and small businesses.[152] Effective solutions that balance the necessity of transparency with the reality of business incentives to protect intellectual investments will be more effective in reinforcing organizational responsibilities while simultaneously delivering useful transparency that satisfy product safety and AI fairness objectives.[153]

---

145. *Id.* at 1362–63.

146. *See* discussion *supra* I.B.1.

147. Daniel Susser, *Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures*, *in* AIES '19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 403, 403–08 (2019).

148. *Id.* at 404.

149. *Id.* at 404–05 (describing technology transparency, or the process by which technology no longer becomes noticeable to users as they focus on tasks to accomplish with the technology rather than the technology itself).

150. *See id.* at 407 (calling for "radical transparency" in sharing information about the technology integrating algorithmic decision-making and increased accountability for algorithmic functionality).

151. Shlomit Yanisky-Ravid & Sean K. Hallisey, *"Equality and Privacy By Design": A New Model of Artificial Intelligence Data Transparency Via Auditing, Certification, and Safe Harbor Regimes*, 46 FORDHAM URB. L. J. 428, 434–35 (2019).

152. *See* Fromer, *supra* note 98, at 717–18, 722–23, 727–28.

153. *See infra* Part IV.

III. FRAMING THE BLACK BOX PROBLEM

A few statutes are have begun mandating explanation of how automated decisions are made.[154] Explanation usually requires both (1) notice of automated decision-making, including how decisions will be made (prospective) and (2) response to automated decision-making, including how decisions were made (responsive).[155] Although it may be possible to communicate a general sense of how decisions are made, dynamic AI decisions are usually rendered in a personalized or highly contextual way, complicating the effectiveness of such explanations.[156]

Further, providing explanation after an algorithm makes a decision may prove impossible or not actually useful. Highly complex algorithms resulting from deep learning systems may be so complex that even the system's creator does not understand how the algorithm made the decision.[157] For less complex systems, although the algorithm may be possible to read, the dynamic nature of the algorithm means that the algorithm making the decision now is not the algorithm that made the previous decision.[158] This means that if explanation is needed, an algorithm must be recorded every time it makes a decision to ensure traceability.[159] While this is possible, it is highly inefficient for dynamic algorithms.[160]

A. AI Necessitates Differentiated Roles and Responsibilities

In addition to calls for validation, actions rejecting secrecy, overall transparency, and due process in general algorithmic decision-making,[161]

---

154. *See* Eduardo Soares, Tariq Ahmed, Ruth Levush, Gustavo Guerra, & James Martin, *Regulation of Artificial Intelligence: The Americas and the Caribbean*, LIBR. OF CONG., https://www.loc.gov/law/help/artificial-intelligence/americas.php#us.

155. *See, e.g.*, GDPR, *supra* note 15. The GDPR addresses explanation both in the ability to receive notice of automated processing or profiling that affects an individual's legal rights and also provides a right of objection, wherein an individual can object to such processing. *Id.*

156. *See supra* Section I.B.

157. *See supra* text accompanying notes 106–20, 128–40.

158. In unsupervised learning, for example, the model is trained on a *dynamic* data set, meaning that the algorithm is changeable over time. Supervised and unsupervised learning models may be complex or simplistic, but unsupervised learning models work on changing data sets, which means that if the data supplied to it materially changes, so does the algorithm associated with it. Ultimately, this means that in order to determine the exact algorithm at the time the decision was made, organizations must maintain a library of historical algorithms, which may be difficult if the data are highly dynamic and the algorithm is prone to frequent change. *See* Marshall Chang, *How is Reinforcement Learning Different from Un/supervised Learning?* INFORMACONNECT (Sept. 8, 2020), https://informaconnect.com/how-is-reinforcement-learning-different-from-unsupervised-learning/.

159. *See generally* DILLON REISMAN, JASON SCHULTZ, KATE CRAWFORD, & MEREDITH WHITTAKER, AI NOW, ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY (2018) (discussing the need for review of the system and potential decisions). The need for reviewable systems presumes access to the algorithm that actually automated the decision. For dynamic systems, this may be tremendously onerous to maintain as well as prone to inaccurate and imprecise recording.

160. *Id.*

161. *See, e.g.*, *Restoring Transparency*, *supra* note 102, at 237–39.

scholars have also called for AI-specific *ex ante* solutions.[162] Jack Balkin, for example, has proposed "the concept of information fiduciaries" for any circumstances involving a disproportionate relationship between an organization and its consumers, including situations involving AI.[163]

Balkin detailed the role of the information fiduciary in three laws of robotics, intended to apply to algorithms from a fiduciary perspective.[164] The first law establishes the role of the primary fiduciary with respect to consumers with comparatively less available information.[165] In part, Balkin associated the degree of black box dealings (the degree to which individuals *must* rely on an organization) as one of the factors in determining whether an organization owes a fiduciary duty.[166] The second law establishes that "algorithmic operators," or organizations employing algorithmic means, owe some duty to the public in addition to existing duties to their customers.[167] Today, manufacturers are accountable to downstream consumers despite the existence of an intermediary, such as a retailer.[168] The third law discusses an obligation to not create a nuisance in the form of undesirable algorithmic traits.[169] Balkin defined nuisance differently than common-law torts as attributes affecting a broad group of people in an undesirable way—for example, for harms that are matters of scale, rather than binary decisions, such as discrimination.[170]

Frank Pasquale responded to Jack Balkin's perspective of information fiduciaries and the cost-benefit determination and identified circumstances like discrimination where trade-offs may not be appropriate, even when they are beneficial for a great deal of other consumers.[171] In his criticism, Pasquale both acknowledged the merits of the first two laws and noted that because the third law with respect to nuisance fails to acknowledge the direct responsibility of an AI operator for harmful AI behavior, it gives operators a "shield of disruptive experimentalism" that

---

162. *See* Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133, 139–42 (2016) (positing that humans should be able to interfere with automated decisioning, but that "AI Guardians," should be created to resolve these potential issues).

163. *See* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1208–09 (2016). *But see* Lina M. Khan & David A. Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV. L. REV. 497, 538–39, 40 (2019) (critiquing the concept of information fiduciaries and positing that perhaps the fiduciary analogy does not appropriately accommodate the realities of dominant online platforms).

164. *See* Balkin, *supra* note 106, at 1217–19 (framing robotics laws within the duties of information fiduciaries).

165. *Id.* at 1228–29.

166. *Id.*

167. *Id.* at 1231–32.

168. *Id.*

169. *Id.* at 1232–33.

170. *Id.* at 1233–35 (citing Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 169 (2017) (discussing the need for algorithmic impact statement on predictive-policing technologies)).

171. *See Toward a Fourth Law of Robotics*, *supra* note 102, at 1248.

"deflect[s] responsibility."[172] Pasquale saw the limitations of cost-benefit economics for AI regulatory enforcement and encouraged "deontological patterns of justification in the technology world to complement the utilitarianism of cost-benefit analyses."[173]

Collectively, Balkin and Pasquale suggested that: (1) additional responsibilities to users should inure when an organization engages in activities that are not visible to consumers; (2) responsibilities also apply with respect to public impact; (3) creation of algorithms should not be a nuisance, which might mean in practice these algorithms are reasonably safe, effective, and fair with respect to consumers or other animate users; and (4) economic cost-benefit analyses should not undermine regulatory regimes simply because they fulfill economic goals.[174]

Overall, these four suggestions are reasonable and compatible with AI technology as long as they are fulfilled using a variety of information sources, such as data selection, infrastructure choices, training data sets and methods, and primary goal identification in human-design choices that affect fairness and safety.[175] Although designed for various types of information fiduciaries, this approach could create clear responsibilities for disclosing comprehensive AI process, system, and infrastructure details, while promoting responsive steps after AI issues are reported.

B. Trade Secrecy Frustrates Direct Transparency Goals

In addition to technology challenges, the natural status of a dynamically inscrutable algorithm as a trade secret frustrates transparency goals that involve "explaining the decision."[176] Academics have proposed a variety of solutions to the algorithmic transparency problem, yet few consider the substantial complexity of AI systems and how this complexity, as well as trade secrecy status, might dually frustrate transparency goals.[177] Scholars like W. Nicholson Price II have prescribed specific sector solutions that require disclosure where regulatory agencies have more preventative control, such as the Food and Drug Administration (FDA), and where such disclosure is, in part, protected through confidentiality commitments.[178] However, Price, Arti Rai, and others have noted that trade secrecy status could frustrate transparency goals, and complex algorithms may even survive reverse engineering—or what I call *natural*

172. *Id.* at 1244–45, 1247–48. In Pasquale's numerous scholarly contributions on this exact topic, a keystone is the concept of responsibility for damaging effects of algorithms and other automated decisional means when the mechanisms for such decisions are opaque.

173. *Id.* at 1250.

174. *See Toward a Fourth Law of Robotics*, *supra* note 102, at 1248–52; Balkin, *supra* note 106, at 1223, 1226–27.

175. *See supra* Sections I.A–B.

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

177. *See supra* Part II; *cf.* Selbst & Powles, *supra* note 102, at 233–34.

178. *See Black-Box Medicine*, *supra* note 102, at 455–57.

trade secrets—dynamically inscrutable algorithms.[179] For algorithms like this, Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker have suggested that vendors should waive trade secrecy protection to enable reviewability or otherwise reduce the efficacy of trade secret protection for public agency AI.[180]

Trade secrecy is an available IP strategy for AI systems where complex AI algorithms will likely be unexplainable and ineligible for alternative IP protection.[181] Price has raised these concerns about algorithmic decision-making and has noted that trade secrecy is often an alternative protective mechanism when algorithms and AI systems are unpatentable.[182] Yet, the problem is even more difficult to solve with modern algorithms: algorithms may be so complex that an AI's creator may not understand how it makes decisions, creating nearly indestructible secrecy.[183] Further, because algorithms are mathematical equations created by a machine, it is unlikely an algorithm will be eligible for patent protection.[184]

Due to their "abstract idea" status, algorithms are not likely candidates for public IP protection, such as patent or copyright,[185] so organizations may have to pursue another strategy for protecting information,

---

179. W. Nicholson Price II & Arti K. Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. 775, 790 (2021).

180. *See* REISMAN ET AL., *supra* note 159, at 14–15. Although this concept may be good in theory, the preeminence of trade secrecy status in the United States makes this unlikely to be successful given that alternative IP protections are unlikely to be available for these technologies.

181. Price II & Rai, *supra* note 179, at 779 (explaining that patents require a level of disclosure that may be difficult to achieve).

182. *See Black-Box Medicine*, *supra* note 102, at 446.

183. *See* Dave Gershgorn, *AI Is Now So Complex its Creators Can't Trust Why it Makes Decisions*, QUARTZ (Dec. 7, 2017), https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/; Marianne Lehnis, *Can We Trust AI if We Don't Know How it Works?* BBC NEWS (June 15, 2018), https://www.bbc.com/news/business-44466213; Victor Tangermann, *This AI Knows When You'll Die and Its Creators Don't Know How*, FUTURISM (Nov. 11, 2019), https://futurism.com/neoscope/ai-knows-when-youll-die-how.

184. *See, e.g.*, *Black-Box Medicine*, *supra* note 102, at 444–46 (highlighting the lack of patent incentives for black-box medicine inherent from its use of algorithms). The Author does not explore the viability of copyright as an alternative due to the idea–expression dichotomy. Copyright has not been particularly useful in other software cases, either, offering at best thin copyright protection.

185. Gottschalk v. Benson, 409 U.S. 63, 67 (1972) (holding that a method for converting binary-code-decimals into binary numbers was invalid as it constituted an "abstract idea" prohibited from obtaining a patent under 15 U.S.C. § 101 subject matter eligibility). Although thin copyright may apply to collections of data, typically an algorithm itself, unless expressive in and of itself rather than a process, will be not be eligible for copyright protection under 17 U.S.C. § 102(b), where "[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work[,]" the idea–expression dichotomy. *See, e.g.*, Experian Info. Sols., Inc. v. Nationwide Mktg. Servs., Inc., 893 F.3d 1176, 1183 (9th Cir. 2018) (quoting 17 U.S.C.S. § 102(b) (LexisNexis 2020)). *But cf.* Lemley & Casey, *supra* note 67, at 105–07 (describing the potential for copyright law to apply to AI technologies as they learn). Trade secrecy, especially for resource-constrained organizations, may be an appealing alternative, especially in light of greater federal enforcement under the Defend Trade Secrets Act (DTSA). *See* Clark D. Asay, *Patenting Elasticities*, 91 S. CAL. L. REV. 1, 26–27 (2017) (describing the use of "substitute" intellectual property protection as an alternative to patent protection).

such as via trade secrecy or other employee-confidentiality obligations.[186] However, secrecy might also stunt innovation in an area positioned to solve the most complex social problems.[187] If information about algorithms is not available, improvements might not develop, and the goal of the patent system may be frustrated.[188] Complete secrecy also removes the possibility of any algorithmic validation, which is crucial to reach both safety and fairness goals.[189] Ultimately, as Price and Rai have observed, "regulatory carrots should come with disclosure obligations," meaning that innovation incentives must nevertheless employ *some* disclosure.[190] Unrestricted innovation can result in over-privatization of information necessary for future innovation, while overly restrictive models also dampen innovative development.

Forced explanation of automated decision-making details may interfere with an organization's reasonable efforts to keep its proprietary information—information that may have significant value—secret, should it opt for trade secret protection.[191] Organizations may resist AI transparency if trade secrecy is the desired approach for an AI system. Due to the inherent necessity of keeping a trade secret—well, secret—explaining how an AI decision is made will likely destroy an algorithm's trade secret status, depending on the extent of the required disclosure.[192]

Trade secrecy operates on a presumption against disclosure and transparency and enables organizations to enforce unpermitted disclosures of organizational trade secrets.[193] While trade secrecy permits disclosure to a limited number of necessary parties within an organization, broad disclosure of trade secrecy destroys trade secret status.[194] For these reasons, organizations that protect inventions using trade-secret status cannot disclose details of the trade secret.[195] However, protecting algorithm-enabled systems via complete trade secrecy (training process, training data set, infrastructure choices, and the algorithms themselves)

---

186. Sharon K. Sandeen, *Through the Looking Glass: Trade Secret Harmonization as a Reflection of U.S. Law*, 25 B.U. J. SCI. & TECH. L. 451, 465–66 (2019).

187. *See* Price II & Rai, *supra* note 179, at 789. In some sectors, such as healthcare, disclosure is central to innovation development. Moreover, innovation policy has established additional "toolkits" beyond traditional notions of intellectual property. *Id.*, at 783.

188. *See* Sandeen, *supra* note 186, at 459–60.

189. *See id.*

190. *See* Price II & Rai, *supra* note 179, at 783.

191. *See infra* Part IV.

192. *See Black-Box Medicine*, *supra* note 102, at 446–48. Of course, not everything kept secret is a trade secret. It must also have independent economic value and reasonable steps must have been taken to maintain its secrecy.

193. *See* Andrea Contigiani, David H. Hsu, & Iwan Barankay, *Trade Secrets and Innovation: Evidence from the "Inevitable Disclosure" Doctrine*, 39 STRAT. MGMT. J. 2921, 2922–24 (2018).

194. *See Black-Box Medicine*, *supra* note 102, at 452–53. Of course, anything protected by a trade secret must not be able to be independently identified, a concept called "reverse engineering." However, for an IT system, it certainly is possible that much of the system could be protected as trade secrets so long as few people have access to each secret and reasonable protocols are used to protect them. *Id.* at 469.

195. *See id.*

means that the entire AI system, not just the algorithm, could be inaccessible to the public.[196]

Effective legal solutions must consider whether a right to explanation and complete transparency of the algorithm itself will actually assuage serious AI concerns or if an alternative might better balance organizational and public interests.[197] If such algorithms and the processes used to create them are indeed protected by trade secrecy status, substantial barriers will exist for transparency advocates; organizations will likely resist sharing algorithms with regulators, the public, or courts.[198]

C. Algorithm, Design, and System Secrecy Are Separate and Distinct Transparency Considerations

Although much academic attention has focused on algorithmic transparency, more attention should be paid to the processes, methods, and strategies used to create algorithmic decision-making systems, as well as the architecture of physical and remote systems controlling the function of these algorithms.[199]

As described in Part I, algorithmic design choices relating to data, training, structure (as in hidden layers), and human validation or feedback loops are critical to creating safe and fair algorithms functioning in the public sphere.[200] As a fundamental question, choices related to the type of AI used, such as machine, neural, or deep learning approaches, and AI's specific tasks and attendant goals are crucial to creating effective AI.[201] The system and its architectural structure also reflect important choices, specifically in relation to AI reliability, availability, and safety concerns.[202] These AI system attributes represent three separate considerations when reviewing AI systems, all of which are relevant for an objective review of AI: (1) prospective choices, or design and system choices made prior to the AI's development, such as system design and development;[203] (2) responsive choices, or changes to these choices made after observing early-stage AI performance (choices made to "improve" the algorithm's outcomes, such as feedback loops that tune the algo-

196. *See id.* at 446–47.

197. *See infra* Part IV.

198. Trade-secret status was strengthened by the Defend Trade Secrets Act (DTSA), a federal law that created a criminal and civil status for trade-secret misappropriation—taking and using (or profiting) from an organization's trade secret. Defend Trade Secrets Act of 2016 § 2, Pub. L. No. 114-153, 130 Stat. 376 (codified as amended at 18 U.S.C. § 1836(b)).

199. *See* Price II & Rai, *supra* note 179, at 777–80. Price and Rai helpfully describe explainability goals as including more than just the algorithm, such as training data, training, process, and model.

200. *Id.* Price and Rai recommend a model that leverages tools to explain system function. Presumably, these tools would probe the system structure of such AI solutions.

201. *See* discussion *supra* Section I.B.

202. *See* discussion *supra* Section I.B.

203. *See* discussion *supra* Section I.B.

rithm);[204] and (3) actual algorithmic function (how the algorithm functions "in the wild," or when it is actively operating for consumers).[205]

It is rare for organizations to routinely maintain an algorithm's secrecy but publicly release the algorithm's underlying design and system architecture. In fact, most for-profit businesses typically aim to protect their AI systems in some way.[206] Although code libraries and open-source codes are popular for traditional coding, AI is comparatively still in its infancy.[207] And when AI inventions become a market differentiator or seek to attract venture capital investment, it is rare that businesses release any specific information about their AI creations publicly.[208]

Certainly, there must be some middle ground for finding an appropriate solution that encourages responsible AI development. Scholars have proposed, for example, limited disclosure to confidential third parties,[209] disclosure related to the process of creating the AI algorithm,[210] disclosure of indicia related to the algorithms,[211] and contrastive or counterfactual explanations.[212] Collectively, these scholars see value in disclosures related to algorithmic functionality in that such disclosures provide a more complete picture of algorithmic decision-making.[213]

---

204. *See* discussion *supra* Section I.B.3.

205. *See* MICHAEL ROVATOS, BRENT MITTELSTADT, & ANSGAR KOENE, CTR. FOR DATA ETHICS & INNOVATION, LANDSCAPE SUMMARY: BIAS IN ALGORITHMIC DECISION MAKING 27 (2019).

206. *See also* W. Michael Schuster, *Artificial Intelligence and Patent Ownership*, 75 WASH. & LEE L. REV. 1945, 1985–91 (2018). Although AI systems are technology-based (and many technologists seek to share information), substantial investment capital has been funneled into AI technologies, which presupposes some ability to maintain IP and confidential information in one way or another.

207. *See* Matthew Russell, *Why You Don't Need to Share Personal Data to Benefit from A.I.*, FORBES (Oct. 3, 2016, 9:00 AM), https://www.forbes.com/sites/forbestechcouncil/2016/10/03/why-you-dont-need-to-share-personal-data-to-benefit-from-ai/?sh=11253f0434c5.

208. *See* Dominic Chalmers, Niall G. MacKenzie, & Sara Carter, *Artificial Intelligence and Entrepreneurship: Implications for Venture Creation in the Fourth Industrial Revolution*, ENTREPRENEURSHIP THEORY AND PRAC. 1, 5–6 (2020).

209. *See Privacy and Accountability*, *supra* note 102, at 12–13 (describing the role of third-party reviewers for medical AI as part of the FDA's clearance processes).

210. *See* Selbst & Barocas, *supra* note 27, at 1133 (proposing that associated documentation of decisions in creating the algorithm be available publicly upon some action, during litigation for example, including an impact statement explaining the relevant choices made by AI designers); Brauneis & Goodman, *supra* note 176, at 131 (discussing value in disclosure of attendant information, including the purpose for which the algorithm was developed and plans for "validation and follow-up.").

211. Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 2, 38 (2019) (describing the sufficiency of descriptions of why automated means may be used and proposing complete openness of algorithms for purposes of testing when used by government agencies).

212. Sandra Wachter, Brent Mittelstadt, & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J. L. & TECH. 842, 845 (2018) (distinguishing between providing explicit explanation in the form of logic description versus providing information that would have led to a different result).

213. *See supra* Section I.B.

IV. FACILITATING AI DISCLOSURE

A limited number of legal mechanisms can functionally protect an algorithm's trade secret while simultaneously promoting disclosure of prospective and responsive choices. This could also enable testing of the live, dynamically inscrutable algorithm. Legal solutions should provide the ability for adverse parties and third parties, rather than only government gatekeepers, to evaluate AI for fairness and safety.[214]

One method to accomplish this is through trusted third parties or certifiers, as suggested by Roger Ford and Price.[215] Price and Alfred Fruh have similarly suggested the possibility of leveraging the patent system as an alternative to trade secrecy in AI,[216] and Price and Rai have suggested the possibility of facilitating information sharing through public funding in data generation or improving disclosure in the patent system.[217] Although several challenges arise with existing patent law doctrine and case law,[218] there are natural benefits to such a model for improving AI information sharing. First, patent law as a disclosure mechanism serves a broader purpose than pure economic incentive.[219] Although patent law may be a poor fit for AI at this time, it is possible to model a disclosure mechanism that offers limited protection while simultaneously promoting disclosure of the system and process information necessary to effectively evaluate AI product safety and fairness.

A. Patent Law's Disclosure History

As Sean Seymore has explained, patent law has a longstanding disclosure requirement which serves to enhance public knowledge and support technological progress.[220] Without disclosure—the quid pro quo for limited exclusivity—the full promise of IP is not realized.[221] In addition to disclosure, the public benefits from any remaining value after patent-term exclusivity has expired.[222] Essentially, this means that the invention previously protected under patent law is now free to the public. It

---

214. *See* David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, & Mariano-Florentino Cuéllar, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 70–78 (2020).

215. *See Privacy and Accountability*, *supra* note 102, at 20.

216. *See* Alfred Frueh, *Transparency in the Patent System – Artificial Intelligence and the Disclosure Requirement*, *in* RETHINKING PATENT LAW AS AN INCENTIVE TO INNOVATION 3 (Rafał Sikorski & Žaneta Zemla-Pacud eds., 2021).

217. *See Black-Box Medicine*, *supra* note 102, at 451.

218. *See* Peter Lee, *Patent Law and the Two Cultures*, 120 YALE L. J. 2, 13, 19, 27, 52 (2010).

219. Sean B. Seymore, *Patenting the Unexplained*, 96 WASH. L. REV. 707, 717 (2019). For example, one well-known alternative interest is the public interest. Patents are available to be used by the public after twenty years exclusivity in most cases. *Id.* This seems to suggest that Congress intended a trade-off: disclosure and later public use to generate more inventions in exchange for legal protection against patent infringement.

220. *Id.*

221. *Id.* at 715; Dmitry Karshtedt, *Did Learned Hand Get It Wrong?: The Questionable Patent Forfeiture Rule of Metallizing Engineering*, 57 VILL. L. REV. 261, 300 (2012).

222. *See* Seymore, *supra* note 219, at 713–14.

is not incompatible to consider that patent law may also serve some additional public benefit in the form of its primary disclosure vehicle. That is, patent law as a disclosure mechanism positioned as an alternative to trade secrecy[223] could be used to promote fairness and safety to the public's benefit.

In addition to benefits of properly implemented disclosure for fairness and safety, patent law also incentivizes innovation.[224] Patent law's functional purpose promotes technological and scientific innovation.[225] Historically, it promotes social benefit, economic benefit, and residual availability after the exclusivity term.[226] Because AI innovation will likely propel sectors forward with significant technical improvements that solve complex problems, it is desirable to incentivize innovation even when inventions may not qualify under current case law.[227] For example, patent availability for diagnostic-imaging AI that can identify Alzheimer's Disease in its very early stages could promote further developments in this important research area in addition to ensuring the AI is nondiscriminatory, reliable, and safe.

In situations where no economic benefit directly incentivizes disclosure, the United States may nevertheless value the properties of disclosure. Indeed, disclosure is the "centerpiece of patent policy" in that "the ultimate goal of the patent system is to bring new designs and technologies into the public domain through disclosure."[228] Specifically, some information may enhance knowledge development in certain fields, where information could otherwise be kept secret.[229] As Sonia Katyal has described, software and computer code is one unique area where patents and trade secrets may be alternative options for protecting an invention; obviously, when natural trade secrets exist, trade secrecy is more attrac-

223. *Cf.* W. Nicholson Price II, *Regulating Secrecy*, 91 WASH. L. REV. 1769, 1779–80 (2016).

224. W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1419–20 (2016) [hereinafter *Big Data*].

225. *See* Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 UNIV. CHI. L. REV. 1017, 1038–39 (1989) (discussing Joseph Schumpter's theory on how patent protections may promote technological innovation).

226. *See* Asay, *supra* note 185, at 24 n.109 (citing John F. Duffy, *Rethinking the Prospect Theory of Patents*, 71 U. CHI. L. REV. 439, 439–40 (2004)).

227. *See Big Data*, *supra* note 224, at 1419. There has been some debate over whether incentivization is necessary for all sectors as AI investment continues to rise. In 2018, IoT and AI led new technology investment areas. Tony Safoian, *AI and IoT Lead the Way for Enterprise New Tech Investment in 2018*, FORBES (May 29, 2018, 7:30 AM), https://www.forbes.com/sites/forbestechcouncil/2018/05/29/ai-and-iot-lead-the-way-for-enterprise-new-tech-investment-in-2018/.

228. *See* Seymore, *supra* note 219, at 713, 715 (emphasis omitted) (quoting Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141, 151 (1989)); *see generally* Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539 (2009) (describing how patent law provides socially desirable benefits both for use after the patent term and as a notice function to stimulate innovation during the patent's term).

229. Seymore, *supra* note 219, at 710, 714.

tive.[230] Although incentives may promote the patent process or alternative public disclosure, without incentives, natural trade secrets will likely become legally enforceable trade secrets.

Alternatively, although patents may be more costly and time-consuming to secure, they also provide stronger protection in the event the disclosed invention has a higher probability of reverse engineering or if the invention cannot be kept reasonably secret.[231] And especially for start-ups or small businesses, a patent may seem more concrete than a secret for would-be investors.[232] Patents, although disfavored for some software inventions, are highly desirable and valuable for inventors and organizations.[233]

B. Current Patentability Restrictions Complicate Patent Enforceability for AI Inventions

Actual AI algorithms are unlikely to be granted patent status under 35 U.S.C. § 101, which bars patenting abstract ideas, including mathematical formulas, as abstract ideas.[234] Moreover, disclosing the details of the algorithm itself through a patent may not be particularly useful as the details are likely to be nonintuitive.[235] However, it may be possible to patent the process for creating the AI algorithm, including the choices that produce an innovative algorithm.[236] Similarly, it may also be possible to patent the system architecture and orientation for supporting algorithmic functionality—or a patent application could include both the process and architecture as part of the overall AI system.[237] However, as Clark Asay and Ryan Abbott have described in great detail, there are serious challenges to patenting the existing AI system,[238] and even when

230. Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1211 (2019).

231. *Id.* at 1215.

232. Although not required to receive venture capital funding, often start-ups are looking for *some* IP strategy to protect the inventions created by a start-up. *Protecting Artificial Intelligence IP: Patents, Trade Secrets, or Copyrights?*, JONES DAY (Jan. 2018), https://www.jonesday.com/en/insights/2018/01/protecting-artificial-intelligence-ip-patents-trad.

233. A recent World Intellectual Property Organization (WIPO) report has noted a staggering increase in AI-related patenting and scientific publishing since 2013, demonstrating a desire *to* patent. *See* WORLD INTELL. PROP. ORG., WIPO TECHNOLOGY TRENDS 2019 ARTIFICIAL INTELLIGENCE 13–14 (2019). Notably, deep learning, a neural-networking AI approach, has increased 175% between 2013 and 2016. *See id.* This is not to say that all organizations and inventors pursue patents but rather to illustrate that there is an appetite to protect AI inventions that appears to be increasing as more inventions are created.

234. *See* Gottschalk v. Benson, 409 U.S. 63, 67–68 (1972). It should be noted that the inexpressive nature of algorithmic equations likely does not make such information a candidate for effective copyright protection either.

235. *See* Selbst & Barocas, *supra* note 27, at 1097–98.

236. Emily J. Gardel & Daniel G. Rudoy, *Strategies for Patenting Artificial Intelligence Innovations in the Life Sciences*, 39 LICENSING J. 1, 1 (2019).

237. *Id.*

238. *See generally* Clark D. Asay, *Artificial Stupidity*, 61 WM. & MARY L. REV. 1187 (2020) [hereinafter *Artificial Stupidity*] (describing the poor fit of patents and trade secrets to protect artificial intelligence); Ryan Abbott, *Everything Is Obvious*, 66 UCLA L. REV. 2 (2019) (arguing that obviousness in AI would be a significant bar for patent eligibility).

patents are granted, they likely lack the disclosive functionality desirable for AI systems.[239]

Existing software-patent practices do not necessarily achieve desirable AI social welfare goals. Often, patents are uninformative for purposes of use or duplication,[240] especially for software patents where even the source code is not supplied.[241] Patentees claim as much as possible for purposes of patent enforcement.[242] In the software industry, often would-be patentees both submit broadly written patent claim language while simultaneously retain as much information as possible for proprietary purposes (including as trade secrets).[243] Ultimately, this means that organizations might both reap the reward of exclusivity while simultaneously benefiting from secrecy.

C. Congress Has Historically Considered Alternatives to Traditional Patent Exclusivity

If no patent protection is available, trade secrecy is a realistic option for organizations.[244] Specifically, for complex AI, the inscrutable nature of the algorithms would make reverse engineering highly unlikely, if not impossible, for AI design choices and system architecture.[245] Like the patent system, trade secrecy offers responsive protection because organizations may recover for trade-secret misappropriation.[246] And increasingly, software and other information technology solutions are protected via trade secret and employee confidentiality.[247] Yet for the reasons previously discussed, trade secrecy does not fulfill any of the important public benefits crucial for resolving AI safety and fairness issues.[248]

With patent law's considerable limitations in mind, precedent exists to create an exclusivity-based model. There are limited circumstances where Congress or administrative agencies have provided patent exclusivity where it otherwise might not be available. For example, the Plant

---

239. *See Artificial Stupidity*, *supra* note 238, at 1208–09.

240. *See* Seymore, *supra* note 219, at 708–09.

241. *See* Katyal, *supra* note 230, at 1213.

242. Although over-claiming can actually be the basis of a criminal case, broadening claims is a common activity that, as Oskar Liivak argues, leads to poor (and not sufficiently disclosive) patents. *See* Greg Reilly, *Guest Post by Greg Reilly: Forcing Patent Applicants to Internalize Costs from Overclaiming*, WRITTEN DESCRIPTION (Nov. 2, 2016), https://writtendescription.blogspot.com/2016/11/.

243. Katyal, *supra* note 230, at 1213 (quoting Greg Vetter, *Are Prior Use Rights Good for Software?*, 23 TEX. INTELL. PROP. L.J. 251, 306 (2015)).

244. Michael O'Brien, *Trade Secrets vs. Patents: Pros & Cons*, O'BRIEN PAT. SOLS. (Aug. 20, 2018), https://www.obrienpatents.com/trade-secrets-vs-patents-pros-cons/#.

245. *See* David V. Sanker, *AI Protection*, INTELL. PROP. MAG., July/Aug. 2020, at 29, https://www.morganlewis.com/-/media/files/publication/outside-publication/article/2020/029-ipm-jul-aug-2020-feat.pdf.

246. *Id.*

247. Michael J. Kasdan, Kevin M. Smith, & Benjamin Daniels, *Trade Secrets: What You Need to Know*, 9 NAT'L L. REV. 346, *1, *4 (2019).

248. *See supra* notes 176–98 and accompanying text.

Patents Act of 1930 provided patent eligibility for inventors who "invent[ed] or discover[ed] and asexually reproduce[ed] any distinct and new variety of plant."[249] Prior to 1930, plant patents were unenforceable due to the § 101 bar on patenting "natural phenomena" and two other factors weighing against traditional patent protection: the inability to adequately describe the plant variety and the fact that plant breeding was not sufficiently reproducible.[250] To permit plant patents and stimulate non-government-funded plant variety development, Congress passed the Plant Patents Act describing patentable subject matter and relaxing the written description requirement.[251] Plant patents enjoy a twenty-year exclusivity period.[252]

In this example, Congress identified a benefit for incentivizing production and improving disclosure. Although there were barriers, especially for subject matter eligibility, Congress saw value in providing legal protection for this type of investment for public benefit.[253] AI may qualify for a similar dispensation. In 2016, President Obama's Administration wrote the first report on AI, entitled "Preparing for the Future of Artificial Intelligence."[254] In it, the National Science and Technology Council highlighted numerous public benefits, yet noted potential discrimination and safety issues.[255] During workshops informing the final report, attendees called for greater transparency but cautioned against focusing on transparency given the complexity of systems, urging more frequent testing rather than transparency.[256] In 2019, President Trump held a White House Summit on AI in Government, and President Trump signed an executive order creating the American AI Initiative.[257] The Executive Order, entitled "Maintaining American Leadership in Artificial Intelligence," references the United States' need to promote AI innovation while also fostering public trust and confidence in AI technologies.[258]

The two administrations' focus on promoting AI technology demonstrates the level of national interest in promoting its development. However, significant concerns related to fairness and safety reduce the speed at which the United States can scale safe, effective, and fair tech-

249. 35 U.S.C.S. § 161 (LexisNexis 2020).

250. OFF. OF TECH. ASSESSMENT, U.S. CONGR., NEW DEVELOPMENTS IN BIOTECHNOLOGY: PATENTING LIFE–SPECIAL REPORT 71 (1989) [hereinafter NEW DEVELOPMENTS IN BIOTECH].

251. *Id.* The creation of the Plant Patents Act also followed similar international developments.

252. U.S. PAT. & TRADEMARK OFF., GENERAL INFORMATION ABOUT 35 U.S.C. 161 PLANT PATENTS, https://www.uspto.gov/patents/basics/types-patent-applications/general-information-about-35-usc-161 (last visited Feb. 25, 2021).

253. NEW DEVELOPMENTS IN BIOTECH, *supra* note 250, at 71.

254. SUBCOMM. ON MACHINE LEARNING AND A.I., EXEC. OFF. OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016) [hereinafter PREPARING FOR THE FUTURE].

255. *Id.* at 13–14, 30–34.

256. *Id.* at 32.

257. *Artificial Intelligence for the American People*, TRUMP WHITE HOUSE ARCHIVES, https://trumpwhitehouse.archives.gov/ai/ (last visited Feb. 26, 2021).

258. Exec. Order No. 13,859, 84 Fed. Reg. 3967 (Feb 11, 2019).

nologies for use by and for U.S. consumers.[259] In order to create the kind of system that promotes basic disclosure and appropriate testing, Congress should consider creating a similar exclusivity to the Plant Patents Act path combined with enhanced disclosure requirements.

D. A Patent-Like Registration System Could Facilitate Effective Disclosure

Why might a patent-based system actually work in this environment?[260] First, a desirable patent-like registration should encourage disclosure while improving fair and safe outcomes through independent evaluation. Further, if AI algorithm hosting and process and system disclosure is mandatory (to meet transparency obligations), offering limited monopoly in exchange for limited disclosure of trade secrets may prevent an unconstitutional taking of IP without compensation.[261] Of course, AI technology inventions carry challenges that are both similar and distinct from software inventions.[262]

In order to create a system that works effectively across multiple industries, a system for AI inventions should address the following areas: (1) the availability of trade secret protection for part of the invention, (2) the exclusivity term, and (3) patent disclosure standards.

First, the availability of trade secrecy for a patent-like protection frustrates patent law's invention disclosure goals, so it is essential that the statute clearly articulates the unavailability of trade secrecy for the scope of what is disclosed in the patent, such as the process or system that creates the algorithm. Although it may then be possible to protect the algorithm itself via trade secrecy, it may be unnecessary for highly complex algorithms as natural trade secrets. The benefit of this approach is that disclosure is incentivized by offering documented legal exclusivity

---

259. *See* PREPARING FOR THE FUTURE, *supra* note 254, at 30.

260. It should be noted that this piece is not intended to discuss why patent may be the only option that works or to discuss whether the patent system is effective. Rather, this approach acknowledges that patent/trade secret implicit trade-off that requires some protection if trade secret is unavailable for at least part of the AI system and process.

261. *See* Ruckelshaus v. Monsanto Co., 467 U.S. 986 (1984). In *Ruckelshaus*, Justice Blackmun observed that "a factor for determining whether a governmental action short of acquisition or destruction of property has gone beyond proper 'regulation' and effects a 'taking' is whether the action interferes with reasonable investment-backed expectations." *Id.* at 987. The destruction of trade secrets for most small businesses, such as start-ups creating AI systems, will likely result in interference with investments based on the technology itself, such as future acquisitions by more established companies. However, trade secret destruction will likely not be considered a taking. As the Court continued, "As long as the appellee is aware of the conditions under which the data are submitted, and the conditions are rationally related to a Governmental interest, a voluntary submission of data in exchange for some economic advantages of a registration can hardly be called a taking." *Id.* (describing whether submission of data to the U.S. Environmental Protection Agency would be considered a taking under that regime).

262. Martin Casado & Matt Bornstein, *The New Business of AI (and How It's Different from Traditional Software*, ANDREESSEN HOROWITZ (Feb. 16, 2020), https://a16z.com/2020/02/16/the-new-business-of-ai-and-how-its-different-from-traditional-software/.

that may be more appealing to potential investors, as well as potentially providing more effective enforcement activities.[263]

Next, AI technologies will likely improve quickly, leaving little remaining public benefit through their unrestricted use. Therefore, a twenty-year exclusivity term is not appropriate in this case. An appropriate period of time should both reflect market dynamics to reserve some residual value to the public while also providing enough protection to incentivize filing. Existing patent exclusivity is currently a one-size-fits-all approach, and some technologies have very little useful public benefit after a patent's expiration.[264] AI's exclusivity time period should take into account incentives for disclosure while retaining some residual public benefit, and the time period specified might be informed by detailed studies to this effect in a variety of industries.

Third, both Congress and the U.S. Patent and Trademark Office (USPTO) must require enhanced disclosure requirements appropriate to the technology. For example, the patent might include either the design approach for producing the algorithm (a process patent) or the system's infrastructure (a system patent). However, the most effective disclosure model would incorporate both patents and disclosure via hosting the live algorithm. Qualifying patents should describe the entirety of the system rather than submitting multiple discrete system component patents. This stringency should correct well-known issues in the existing patent system around information technology products.[265]

AI "patents" must be drawn to the creation of an algorithm, and the purpose of the algorithm should be explicit. Then, organizations seeking patents must provide information about how they plan to use the AI system, which would include its potential risks and benefits, and details about its construction, as Andrew Selbst and Solon Barocas suggest.[266] This approach should provide details about the social context and overall functional purpose of the AI, constraining how the AI is vetted and evaluated by outsiders, similar to product labeling.

---

263. The degree of enforceability for patent infringement in unknown; however, sometimes the threat of infringement is enough to stymie certain competition.

264. Lucas S. Osborn, Joshua M. Pearce, & Amberlee Haselhuhn, *A Case for Weakening Patent Rights*, 89 ST. JOHN'S L. REV. 1185, 1240–42 (2015).

265. Creating greater stringency around patent claiming in software or software-adjacent technologies like AI should prevent overly broad functional claiming that has created the problems illustrated by *Alice v. Kappos*. *See* Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 WIS. L. REV. 905, 962–63 (2013).

266. *See* Selbst & Barocas, *supra* note 27, at 1133–35; *see also* REISMAN ET AL., *supra* note 159, at 15–20 (proposing the use of Algorithmic Impact Assessments (AIAs), wherein potential risks could be identified prior to AI use). AIAs could easily be disclosed by the AI managing company and hosted on the site where the live algorithm is or otherwise added later when third parties conduct these assessments. Algorithmic hosting and disclosure would promote the kind of inspection advocates and researchers are seeking.

In addition to the written documentation meeting an enhanced degree of disclosure both for public benefit and for solidity of the patent itself, Congress should require a mandatory link, hosted by the USPTO, to a user interface with an active version of the live algorithm. The benefit of having a mandatory link to the live algorithm is that the USPTO validates that the algorithm works for purpose of enablement as well as providing a live link for testing and validation.[267] It could also host public, visible feedback on the algorithm, such as fairness or discrimination assessments or ratings, or safety results.

Further, as a condition of patent exclusivity, inventors must maintain an active link to the live algorithm throughout its life, even after exclusivity has expired, and create archival copies of point-in-time algorithms upon material change, or at least monthly. This arrangement supports a known software-testing method called "black-box testing," which evaluates the functionality of the black box through submitting inputs and evaluating outputs.[268] Although black-box testing does not provide complete transparency, it permits testers to see patterns by feeding in a variety of data points and viewing the outputs.[269] Because algorithms may be dynamically inscrutable, there is low risk for reverse engineering, minimizing any likelihood of trade secret destruction.[270]

The benefit to such a hosting situation, as Talia Gillis and Jann Spiess have suggested, allows for "discrimination stress test[ing]," amongst other types of testing, such as testing to identify potential bugs that could reduce consumer safety.[271] Hosting publicly enables adverse parties such as competitors, government entities, advocacy groups, and other experts to actively test the inscrutable algorithms and flag potential issues, which could even be listed on the hosted page or require responses pursuant to an administrative agency's purview (such as the FDA for medical AI).[272] In the event the AI system is adapted to a new use, such as a new self-driving car, a new municipality's public water system, or a new medical device, these new uses and associated context should be updated on the test page.

The benefit of live hosting and testing is that nonlegal processes could easily dovetail for the benefit of consumers. For example, an independent rating system could rate the potential for discrimination or safety issues, similar to Consumer Reports. Alternatively, the government could

---

267. Of course, a cost will be assigned for hosting purposes. Likely, this could be a portion of the AI patent application fee.

268. *Black Box Testing*, SOFTWARE TESTING FUNDAMENTALS (Sept. 17, 2020), http://softwaretestingfundamentals.com/black-box-testing/.

269. *Id.*

270. *See id.*

271. Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 UNIV. CHI. L. REV. 459, 484 (2019) (describing a discrimination stress test that would be conducted after a "supervisory model" is made public, but before public consumers are affected by it).

272. *See id*.

revoke exclusivity in the event an AI system is determined to be unfair or unsafe until the AI system can be retested and demonstrated to be safe and fair. An independent AI oversight agency could manage this mechanism.

E. A Patent Registration-Hosting System Does Not Eliminate the Potential for More Stringent State or Federal Government Mandates

There are a variety of circumstances where additional legal requirements might be desirable due to the risk of the AI's use or the potential discriminatory impact on a population—for example, as the GDPR has noted, when AI use impacts individual rights.[273] A registration-hosting system can complement these processes. For example, an FDA medical device approval process could formally take comments and reviews into account in their evaluation, or the Equal Employment Opportunity Commission could include evaluations when investigating the use of AI interviewers to select candidates for employment.

In limited cases, explanation may be necessary. For example, for predetermined uses, especially when the government is using AI to make determinations about its citizens, such as eligibility for public support, prison time, or responding to legitimate due process concerns.[274] For these situations, additional information may be needed to make the explanation more generally understandable by people with less algorithmic expertise. [275] Although the need for explanation should be limited to specific cases, a registration and hosting solution would supplement these explanations. Moreover, the hosting database with associated feedback could also be statutorily mandated in situations where the probability of negative impact to individuals is high. For both average consumer algorithms and high-impact algorithms, hosted "versions" could be captured at key time-stamped moments, enabling more effective due process and other legal actions.

---

273. *See, e.g.*, GDPR, *supra* note 15, at recital 32 (requiring that consent be given before an individual's personal data may be processed by a website).

274. *See Technological Due Process*, *supra* note 102, at 1249–50 (introducing a system of "technological due process" and arguing it is necessary in order to carry out the procedural protections created in 20th century jurisprudence); *Scored Society*, *supra* note 102, at 1 (arguing that the due process should inform safeguards where individuals are "scored" based on credit and other data); Crawford & Schultz, *supra* note 83, at 93 (proposing a "procedural data due process" right); *see also* Brauneis & Goodman, *supra* note 176, at 128–29.

275. *See Technological Due Process*, *supra* note 102, at 1249–50 (introducing a system of "technological due process" and arguing it is necessary in order to carry out the procedural protections created in twentieth century jurisprudence); *Scored Society*, *supra* note 102, at 1 (arguing that the due process should inform safeguards where individuals are "scored" based on credit and other data); Crawford & Schultz, *supra* note 83, at 93 (proposing a "procedural data due process" right); *see also* Brauneis & Goodman, *supra* note 176, at 128–29.

CONCLUSION

The deafening call for Al transparency in international law and from U.S. scholars demonstrates why it is important to provide more information and access to AI systems. The potential risks, including discrimination and safety issues, may cause not only individual mistreatment or unconstitutional behavior but also physical health risks and property damage.[276] Without the appropriate system to best leverage collective expertise and ensure that integral information is shared, it is unlikely that the public will benefit from fair, safe AI systems.

AI is increasingly inscrutable, sometimes dynamically inscrutable, yet the underlying design choices and system architectures have a significant bearing on an AI algorithm's functionality. Pursuing an alternative to traditional patents—modeled after disclosure concepts that have historically promoted disclosure—will provide effective supporting documentation and an accountable, trust-based system for hosting the live algorithm. Availability of the current, live algorithm for testing purposes provides opportunities for collective feedback and subsequent improvement while allowing companies to simultaneously benefit from limited exclusivity.

The United States should employ solutions that effectively balance business interests and social need for more, better, and safer AI with the public need to promote social welfare through fair, non-discriminatory automated decisioning. The U.S. government can promote AI system and process disclosure through an alternative, parallel patent path; permit the retention of natural trade secrets in the inscrutable algorithm itself; and host the algorithm along with attendant testing outcomes and risk assessment information. This proposal, though audacious, may encourage broad disclosure, even for AI systems that might seem innocuous on the surface, while navigating the challenges of algorithmic complexity through calls for greater transparency.

276. *See, e.g.*, Michael Brenner, Jeannie Suk Gersen, Michael Haley, Matthew Lin, Amil Merchant, Richard Jagdishwar Millett, Suproteem K. Sarkar, & Drew Wegner, *Constitutional Dimensions of Predictive Algorithms in Criminal Justice*, 55 HARV. C.R.-C.L. L. REV. 267, 274 (2020).