Fordham Urban Law Journal

Volume 46
Number 2 *Artificial Intelligence and Predictive Algorithms: Why Big Data Can Lead To Big Problems*
Article 5

2019

"Equality and Privacy by Design": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes

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

Shlomit Yanisky-Ravid & Sean K. Hallisey[*]

ABSTRACT

Artificial Intelligence and Machine Learning (AI) are often described as technological breakthroughs that will completely transform our society and economy. AI systems have been implemented everywhere, from medicine, transportation, finance, art, to legal and social spheres, and even in weapons development. In many sectors, AI systems have already started making decisions previously made by humans. Promising as AI systems may be, they also pose urgent challenges to our everyday life. While much attention has concerned AI's legal implications, the literature suffers from a lack of solutions that account for both legal and engineering practices and constraints. This leaves technology firms without

---

[*] Professor Shlomit Yanisky-Ravid, Ph.D., Fordham Law School, Visiting Professor; Fordham Law Center on Law and Information Policy (CLIP), Head of AI-IP and Blockchain Project; Yale Law School, Information Society Project (ISP), Fellow; Ono Law School, Israel, Senior Faculty, the Shalom Comparative Legal Research Institute, OAC, Founder and Academic Director. Sean K. Hallisey, Fordham Law, CLIP, AI-IP Project, Fellow. We gratefully dedicate this Article to Joel Reidenberg, the founder and the head of Fordham Law Center of Law and Information Policy (CLIP), for his initiative, support, and encouragement, all of which tremendously contributed to the writing of this Article and the development of its ideas. We would also like to thank all the Fellows at the Fordham CLIP IP-AI and Blockchain Project, Yale Law, ISP, as well as to the students of the course "Intellectual Property and the Challenges of Advanced Technology: AI and Blockchain," for their wonderful discussions, insights and comments. Finally, we thank Dean Matthew Diller, Fordham Law School, for promoting and stressing the challenges of advanced technology, data privacy, and intellectual property, and Linda Sugin, Associate Dean for Academic Affairs at Fordham Law, for her support.

guidelines and increases the risk of societal harm. It also means that policymakers and judges operate without a regulatory regime to turn to when addressing these novel and unpredictable outcomes. This Article tries to fill the void by focusing on data rather than on the software and programmers. It suggests a new model that stems from a recognition of the significant role that the data plays in the development and functioning of AI systems.

Data is the most important aspect of teaching AI systems to operate. AI algorithms begin with a massive preexisting dataset, which data providers use to train the system. But the data that AI systems "swallow" can be illegal, discriminatory, altered, unreliable, or simply incomplete. Thus, the more data fed to the AI systems, the higher the likelihood that they could produce biased, discriminatory decisions and violate privacy rights. The Article discusses how discrimination can arise, even inadvertently, from the operation of "trusted" and "objective" AI systems.

To address this problem, this Article proposes a new AI Data Transparency Model that focuses on disclosure of data rather than, as some scholars argue, focusing on the initial software program and programmers. The Model includes an auditing regime and a certification program, run either by a governmental body or, in the absence of such entity, by private institutions. This Model will encourage the industry to take proactive steps to ensure and publicize that datasets are trustworthy. The suggested Model includes a safe harbor, which incentivizes firms to implement transparency recommendations even without massive regulatory oversight. From an engineering point of view, the Model recognizes data providers and big data as the most important components in the process of creating, training and operating AI systems. Even more importantly, the Model is technologically feasible because data can be easily absorbed and kept by a technological tool. Further, this Model is also practically feasible because it follows already existing legal frameworks of data transparency, such as the ones being implemented by the FDA and the SEC.

Improving transparency in data systems would result in less harmful AI systems, better protect societal rights and norms, and produce improved outcomes in this emerging field, especially for minority communities that often lack resources or representation to challenge AI systems. Increased transparency of the data used while developing, training or operating AI systems would mitigate and reduce these harms. Additionally, to better identify the risks of faulty data, industry players must conduct critical evaluations and audits of the data used to train AI systems; one way to incentivize this is a

certification system to publicize good-faith efforts to reduce the possibility of discriminatory outcomes and privacy violations in AI systems. This Article strives to incentivize the creation of new standards, which the industry could implement from the genesis of AI systems to mitigate the possibility of harm, rather than post-hoc assignments of liability.

TABLE OF CONTENTS

INTRODUCTION

Commentators and experts frequently herald artificial intelligence ("AI") as a technological breakthrough that will completely transform our society and economy.[1] From medicine to transportation, finance to art, legal systems to social structures, and many other sectors, AI systems hire, fire, grant loans, predict diseases, and decide who will go to jail and how long they will stay there.[2] Many decisions previously determined by humans are now made by autonomous AI systems.[3] These AI systems, embedded in computers

---

1. *See* MCKINSEY GLOBAL INST., ARTIFICIAL INTELLIGENCE: THE NEXT DIGITAL FRONTIER? 4 (2017), https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx [https://perma.cc/JF98-XLCJ] ("Artificial intelligence is poised to unleash the next wave of digital disruption, and companies should prepare for it now. We already see real-life benefits for a few early-adopting firms, making it more urgent than ever for others to accelerate their digital transformations. . . . AI investment is growing fast, dominated by digital giants such as Google and Baidu. Globally, we estimate tech giants spent $20 billion to $30 billion on AI in 2016, with 90 percent of this spent on R&D and deployment. . . . [E]arly AI adopters that combine strong digital capability with proactive strategies have higher profit margins and expect the performance gap with other firms to widen in the future . . . . [E]arly adopters are already creating competitive advantages, and the gap with the laggards looks set to grow. A successful program requires firms to address many elements of a digital and analytics transformation, including set up the right data ecosystem.").

2. *See, e.g.*, Hilke Schellmann & Jason Bellini, *Artificial Intelligence: The Robots Are Hiring*, WALL ST. J. (Sept. 20, 2018), https://www.wsj.com/articles/artificial-intelligence-the-robots-are-now-hiring-moving-upstream-1537435820 [https://perma.cc/X26T-556Z]; Ingrid Lunden, *Kabbage Gets $200M from Credit Suisse to Expand Its AI-Based Business Loans*, TECH CRUNCH (Nov. 11, 2017), https://techcrunch.com/2017/11/16/kabbage-gets-200m-from-credit-suisse-to-expand-its-ai-based-business-loans/[https://perma.cc/HV64-EWN6]; Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html [https://perma.cc/MD23-6P4R]; Steve Lohr, *IBM Creates Watson Health to Analyze Medical Data*, N.Y. TIMES (Apr. 13, 2015), https://bits.blogs.nytimes.com/2015/04/13/ibm-creates-watson-health-to-analyze-medical-data/ [https://perma.cc/J7JV-DPCH] (describing the practical applications of AI to various industries).

3. *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 633 (2017) ("The accountability mechanisms and legal standards that govern such

and robots, have begun to automate workplaces and have created new applications that rely on the vast amounts of data produced by society's daily occurrences.[4] Corporations, governments, and individuals are investing in the AI sector, creating the specter of a new Industrial Revolution. But if society comes to over-rely on AI too rapidly, it risks overlooking potential problems that may arise.[5] It is true that machine learning offers broad opportunities for

---

decision processes have not kept pace with technology. The tools currently available to policymakers, legislators, and courts were developed to oversee human decisionmakers and often fail when applied to computers instead. For example, how do you judge the intent of a piece of software? Because automated decision systems can return potentially incorrect, unjustified, or unfair results, additional approaches are needed to make such systems accountable and governable."); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RES. 1, 1 (2018) ("Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks."); HANNAH FRY, HELLO WORLD: BEING HUMAN IN THE AGE OF ALGORITHM 25–48, 49–78, 141–74 (2018) (discussing the biases and risks of AI systems in different fields, such as justice, data analytics and crime, in contrast to the trust and faith the public give to advanced technology); *see also* MCKINSEY GLOBAL INST., *supra* note 1, at 31–69 (describing the use of AI systems and its implications in different fields, including retail, electric utility, manufacturing, healthcare and education); Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 5–6 (2014) (discussing the vast use of AI systems versus its risks); Hilke Schellmann & Jason Bellini, *Artificial Intelligence: The Robots Are Now Hiring — Moving Upstream*, WALL ST. J. (Sept. 20, 2018), https://www.wsj.com/video/series/moving-upstream/artificial-intelligence-the-robots-are-now-hiring-moving-upstream/2790C6B9-4E47-4544-9331-36DB418366CF?mod=searchresults&page=1&pos=3 [https://perma.cc/V4XV-RX9A] (noting that "nearly all Fortune 500 companies" are using tools that deploy AI to weed out job applicants, including a video that discusses biases and fairness); DELOITTE, THE STATE OF THE DEAL: M&A TRENDS 2019 (2018) https://www2.deloitte.com/content/dam/Deloitte/us/Documents/mergers-acqisitions/us-mergers-acquisitions-trends-2019-report.pdf [https://perma.cc/D2L2-C2YK] (detailing 1,000 executives' expectations and involvement with M&A activity and emerging technologies); Leon Saunders Calvert, *Using AI to Predict Opportunity in M&A*, REFINITIV: DEAL INSIGHTS (June 20, 2018), https://blogs.thomsonreuters.com/financial-risk/ai-digitalization/using-ai-to-predict-opportunity-in-m-and-a/ [https://perma.cc/TXU4-4SA8].

4. *See* Jason Bellini, *The Robot Revolution: The New Age of Manufacturing*, WALL ST. J. (Feb. 1, 2008), https://www.wsj.com/video/series/moving-upstream/the-robot-revolution-the-new-age-of-manufacturing-moving-upstream/0C3B7686-7D97-4BCE-980B-FAED24F27672 [https://perma.cc/TY36-6RGT] (reporting that hundreds of millions of jobs are affected by AI, and trillions of dollars of wealth are created by replacing employees).

5. *See* Kai-Fu Lee, *The Human Promise of the AI Revolution*, WALL ST. J. (Sept. 14, 2018), https://www.wsj.com/articles/the-human-promise-of-the-ai-revolution-1536935115 [https://perma.cc/UEC2-E4QV] ("The AI revolution will be of the magnitude of the Industrial Revolution — but probably larger and definitely faster.").

innovation in a host of areas such as climate and physical, transactional, and behavioral data about people, pandemics, pharmaceuticals, infrastructure, and supply chains.[6] However, as AI technologies grow in prominence and become more easily implementable, stakeholders must acknowledge that AI has the dangerous potential to violate laws and societal norms.[7]

The growing AI industry is dominated by huge firms that collect, hold, or can afford to access massive amounts of data.[8] But data can be flawed — indeed, instances abound of massive companies utilizing AI systems that produce biased outcomes. In one instance, Amazon's AI facial recognition software, Rekognition, wrongly identified twenty-eight members of Congress as individuals who had jail mugshots.[9] These results demonstrate the existence of race and gender biases present in facial recognition AI system.[10] Similarly, Facebook's software is known to identify the "ethnic affinities" of users' characteristics, which advertisers can then use to exclude

---

6. *See* Julie E. Cohen, *What Privacy Is for*, 126 HARV. L. REV. 1904, 1921–22 (2013).

7. *See* MCKINSEY GLOBAL INST., *supra* note 1, at 4, 8 (stating that "AI promises benefits, but also poses urgent challenges that cut across firms, developers, government, and workers," and that "machine learning has limitations. For example, because the systems are trained on specific data sets, they can be susceptible to bias; to avoid this, users must be sure to train them with comprehensive data set."); Liam Hanel, *A List of Artificial Intelligence Tools You Can Use Today — For Businesses*, MEDIUM (July 11, 2017), https://medium.com/@LiamHanel/a-list-of-artificial-intelligence-tools-you-can-use-today-for-businesses-2-3-eea3ac374835 [https://perma.cc/KB54-S3YE]; *see also* Daniel Newman, *AI and ML Prediction to 2019*, FORBES (July 23, 2018, 10:56 AM), https://www.forbes.com/sites/danielnewman/2018/07/23/three-ai-and-machine-learning-predictions-for-2019/#238d7c784948 [https://perma.cc/7HHX-K96F].

8. *See* MCKINSEY GLOBAL INST., *supra* note 1, at 14; *see also* Rana el Kaliouby, *This App Knows How You Feel from the Look on Your Face*, TED (May 2015), https://www.ted.com/talks/rana_el_kaliouby_this_app_knows_how_you_feel_from_the_look_on_your_face/transcript?language=en [https://perma.cc/MDM5-8DAU] (describing how data was collected from 2.9 million face videos for a project that began at MIT).

9. Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, ACLU (July 26, 2018, 8:00 AM), https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 [https://perma.cc/XZ4K-EZXG] ("Nearly 40 percent of Rekognition's false matches in our test were of people of color, even though they make up only 20 percent of Congress.").

10. Buolamwini & Gebru, *supra* note 3, at 1 (demonstrating empirically that AI systems can discriminate based on classes like race and gender, and evaluating the biases present in automated facial analysis algorithms and datasets by percentage with respect to phenotypic subgroups).

certain users from viewing particular promotions.[11] These troubling, biased consequences are not inevitable in an era of Autonomous, Automated, and Advanced AI Systems — the so-called "3A Era." Rather, they highlight that AI technologies pose crucial challenges that policymakers must address.[12] These challenges cut across firms, developers, governments, and employees; therefore, proper legal and regulatory schemes must be established to ensure that AI development is neither held back nor goes too far.[13]

The innovations in AI technology are moving too fast for Congress to effectively understand and grapple with. Inadequate regulatory schemes might be unbalanced: too permissive a scheme would give cover to and perpetuate existing discrimination in AI programs, while a scheme too restrictive would halt the development of AI technology altogether, stymying its potential benefits. Thus, it is vital to create a framework that can help the industry, the public and policymakers identify where problems with data occur, how they occur, and why they occur. Once these nuances are better understood, the government can more effectively regulate the AI industry. To that end, this Article proposes an AI Data Transparency Model that focuses on illuminating how AI systems utilize *data*. This Model differs from other commentaries on the risks of AI systems in that it does not oppose the use or expansion of AI systems. Rather, this Model recognizes that regulatory schemes have to focus on the *source* of threats and hazards in AI systems — the data itself.

The Transparency Model recommends an auditing and certification regime that will encourage transparency, and help developers and individuals learn about the potential threats of AI, discrimination, and the continued weakening of societal expectations of privacy. If firms choose to utilize non-infringing data from beginning to end,

---

11. Julia Angwin & Terry Parris Jr., *Facebook Lets Advertisers Exclude Users by Race*, PROPUBLICA (Oct. 28, 2016, 1:00 PM), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race [https://perma.cc/5GGJ-4SNK] ("The ubiquitous social network not only allows advertisers to target users by their interests or background, it also gives advertisers the ability to exclude specific groups it calls 'Ethnic Affinities.' Ads that exclude people based on race, gender and other sensitive factors are prohibited by federal law in housing and employment.").

12. *See generally* Shlomit Yanisky-Ravid & Xiaoqiong (Jackie) Liu, *When Artificial Intelligence Systems Produce Inventions: The 3A Era and an Alternative Model for Patent Law*, 39 CARDOZO L. REV. 2215 (2018) (discussing why intellectual property laws have become irrelevant, outdated and inapplicable in the 3A Era, when AI systems produce patentable inventions or copyrightable works of art, and suggesting an alternative model for patent law).

13. MCKINSEY GLOBAL INST., *supra* note 1, at 4.

from the very first steps of developing and training AI systems through the actual operation of those systems, the likelihood of discriminatory outcomes and privacy violations will be greatly reduced.

The proposed Transparency Model takes into account the nature of how AI systems work and the prevalence of multiple stakeholders, each of whom is responsible for developing and operating AI systems (the "Multi-Player Model").[14] These stakeholders may include software programmers, data providers, users, sellers and distributors of AI systems, manufacturers, and others such as the public and the shareholders of firms.[15] As part of the regulatory scheme of the Model, we first contend that each of these stakeholders, especially the data providers, should concern themselves with potential adverse outcomes that AI systems might create. Stakeholders must consider the possibility that AI systems will misinterpret data and produce discriminatory outcomes or otherwise violate human rights. The Model includes a certification process, whereby stakeholders can align, assert, and publicize their efforts to produce AI systems that conform with a transparency industry standard. This certification can be determined internally, or conducted by a third-party auditing agency; either way, the purpose is to encourage the development of a certifiable, uniform industry standard. This Article argues that cultivation of a strong certification process is soundly justified by law and economics and would spur public demand for ethical use of AI. Finally, the Model will raise awareness about the dangers that may arise when stakeholders overlook the possibility that certain compositions of data can have discriminatory effects.

Just as technology has exploded in the 3A Era, so has the literature concerning the legal implications of AI's proliferation.[16] However, the literature to date has tended to focus on the operation of AI

---

14. Yanisky-Ravid & Liu, *supra* note 12, at 2231–36 (coining the "Multi-Player Model" and describing the affiliation between each of the entities in AI systems and the challenge to ownership and accountability that a model of many stakeholders imposes on the AI industry). The development of AI systems is a multi-faceted process, involving numerous "stakeholders." Such stakeholders include data collectors, data aggregators, programmers, trainers, operators, all the way up to executives who market and sell AI services. We use the term "stakeholders" to encapsulate each of these roles, because we argue that each of these distinct actors should concern themselves with the ramifications of AI systems and grapple with their effects.

15. *See id.*

16. *See* Ryan Calo, *Artificial Intelligence Policy: A Primer and a Roadmap*, 51 U.C. DAVIS L. REV 399, 401, 403 (2017).

systems, rather than on the data used to train them.[17] This leaves technology firms without guidelines, which increases the risk of societal harm and leaves policymakers and judges without a regulatory regime to turn to when addressing the novel and unpredictable outcomes of AI systems. This Article also tries to fill that void with the Transparency Model, which focuses on data, rather than on software programmers or algorithms. It is important to mention that some scholars, notably Professor Joel Reidenberg, have opposed the prevailing position that transparency of software and algorithms alone will not completely resolve existing issues of bias and prejudice in AI.[18] This work goes a step further to suggest great emphasis and focus must be placed on the data itself. Focusing on the data is vital and can usher in newfound understanding of how and to what extent AI systems should be integrated into nearly any aspect of society.

A short review of some other important works that have been conducted regarding AI systems demonstrates that a thorough discussion of data itself is missing from the literature at large. In a landmark article, Solon Barocas and Andrew D. Selbst, two scholars who have pioneered the study of the effects of big data and the advent of the internet on individuals' privacy and civil rights, noted that algorithms and the use of big data compromise the spirit of decades old anti-discrimination statutes.[19] They warned of the need to pass new statutes that counteract the dangers that algorithms and big data pose to society.[20] There has been a steady creep of AI into consumer finance, and it remains an unresolved question who should bear the burden of ensuring that their AI applications do not discriminate, target, or fail to provide services to protected demographics.[21] Current laws are insufficient to address these risks, but overregulation could hinder the development of the technology.[22] Many others have raised concerns about the new challenges that AI

---

17. *See id.* at 402 (collecting examples of scholars raising issues pertaining to the effects of AI systems and the "vast increase in computational power and access to training data . . . ").

18. *See* Kroll et al., *supra* note 3, at 658 ("However, transparency [of source code] alone is not sufficient to provide accountability in all cases.").

19. *See generally* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

20. *Id.* at 671 ("[A]n algorithm is only as good as the data it works with.").

21. *See, e.g.*, Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148 (2016) (arguing that lenders should bear that burden).

22. *See id.* at 189–90.

systems pose for criminal justice. For example, scholars have discussed the inadequacy of current legal doctrines in protecting citizens from automated suspicion algorithms, which identify suspects and suspicious activity that would ordinarily be identified by a human police officer.[23]

There are numerous problematic features of machine learning algorithms that make regulation difficult.[24] First, there is *discreetness*, the fact that machine learning applications can be developed with limited visible infrastructure.[25] Next, because so many different entities develop machine learning applications, *diffuseness* makes it difficult to identify who should be regulated.[26] Further, the *opacity* of the developing process creates the possibility that the machine learning application will produce outcomes that are not traceable to particular inputs, and it might be difficult, if not impossible, to retroactively determine the rationale of the decision.[27]

Policymakers cannot hope to resolve all the issues identified above by increasing transparency alone, but resolution certainly requires transparency. Consider discreetness, for example: with better transparency, it would be possible to account for the infrastructure that is present. Creating a log or record to detail who worked on a particular application and what their work entailed is certainly possible.[28] Additionally, making the data sources of an AI application transparent would help control outcomes by ensuring that the application is using the "right" data, rather than impermissible ones. Increasing transparency, thus, will contribute to a larger consensus: something needs to be done to address the new challenges created by AI and machine learning systems.

This Article discusses how workflows and bottlenecks in AI development illuminate policy responses that could reveal the data sources used to train AI systems. It further identifies data issues that

---

23. *See* MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION 28, 96 (2018); *see also* GABRIEL HALLEVY, WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW 16, 21 (2013) ("Some researchers argue that the current law is inadequate for dealing with AI technology and that it is necessary to develop a new legal domain called *Robot Law*.").

24. *See* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 369 (2016).

25. *Id.* at 369–70.

26. *Id.* at 370.

27. *Id.* at 371, 373.

28. *See* Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 8 (2017).

policymakers should focus on and proposes a Model of Data Transparency that could solve some of these issues. Finally, this Article puts forth three modest recommendations. First, stakeholders in the development of AI systems should take steps to audit the data used to train AI systems in order to ensure that the data does not violate regulatory requirements surrounding discrimination and privacy, or the rights of copyright holders. Second, a certification process will help consumers and policymakers understand what is at stake and will clearly indicate to the public what operators ought to ensure that data is used properly. Finally, a safe harbor approach that would limit liability for AI operators in certain circumstances, such as where an AI operator takes significant effort to avoid data misuse but a transgression occurs nonetheless.

This Article proceeds in three parts. Part I briefly describes how AI systems operate and how they develop, focusing on the important role that data plays in these processes. Next, Part II examines issues surrounding datasets that can discriminate or violate privacy norms. Part III then presents the AI Data Transparency Model and discusses how it will help ameliorate the issues examined in Part II. The main objective of this Article is to highlight that AI systems are not free of faults and vices, and to stress the vital role of data transparency in addressing these problems. Through this framework, policymakers can better understand the interests at stake with respect to AI systems and the role they play in society, the economy, and the world.

I. DATA MATTERS: TRAINING THE AI

AI systems are different from traditional algorithms in that they incorporate human-like thought processes that enable them to make decisions autonomously.[29] Throughout the development of AI systems, many different stakeholders offer critical contributions. One of the most important phases of creating AI systems is "teaching" them to operate, which starts with a preexisting dataset that data providers use to train the systems.[30] These providers can be programmers, trainers, the ones who enable access to data, or the

29. *See* Shlomit Yanisky-Ravid, *Generating Rembrandt*, 2017 MICH. ST. L. REV. 659, 661–63 (2017); *see also* Hurley & Adebayo, *supra* note 21, at 159 (defining an algorithm as "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as an output") (quoting THOMAS H. CORMEN ET AL., INTRODUCTION TO ALGORITHMS 1 (3d ed. 2009)).

30. *See, e.g.*, *Training ML Models*, AMAZON MACHINE LEARNING, https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html [https://perma.cc/YN3A-F9TX].

systems' users — whatever entity assembles the data is, in a sense, the data provider. By studying the data, the AI system learns to recognize patterns and similarities; as the system absorbs more datapoints, its capabilities grow in an evolving and never-ending process.[31] Whereas an algorithm or formula creates outputs that derive from fixed weights attached to input variables, an AI system adjusts its weights according to the patterns it identifies from ideal outcomes chosen by the data provider.[32] Even with this control over the data provided, AI systems often remain black boxes: They may be able to correctly and consistently predict a particular outcome, such as the likelihood of credit default, but they cannot explain the *reasons* for this conclusion.[33]

AI systems have become ubiquitous and easy to develop. The proliferation of AI systems has resulted in numerous ready-to-use options available for free download on the internet.[34] Once the AI structure is installed, the data trainer exposes it to vast amounts of data, teaching it what outcomes are desirable and to reject unwanted ones. Through this process, the AI learns to recognize patterns that could lead it to identify positive matches on its own.[35] The current explosion of AI applications would not have been possible without the advent of "Big Data" and the ability of entities to collect massive amounts of information. Huge repositories of data exist on the internet,[36] but trainers can also create their own datasets or rely on third parties to collect new data.[37] However, a critical limitation of

---

31. Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 189–90 (2017); *see also* Yanisky-Ravid, *supra* note 29, at 672–81(describing the process of developing AI systems and their ten human-like features, such as creativity, autonomy and unpredictability).

32. *See* Lee Bell, *Machine Learning Versus AI: What's the Difference?*, WIRED (Dec. 1, 2016), https://www.wired.co.uk/article/machine-learning-ai-explained [https://perma.cc/8Y3K-5KTG].

33. 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/ [https://perma.cc/S52Q-CWHA].

34. *See, e.g.*, TENSORFLOW, https://www.tensorflow.org/ [https://perma.cc/MPN7-SNLM].

35. Hence, the advent of the "Big Data" era. Merriam-Webster defines "big data" as "an accumulation of data that is too large and complex for processing by traditional database management tools." *Big Data*, MERRIAM-WEBSTER (Online ed. 2019), https://www.merriam-webster.com/dictionary/big%20data [https://perma.cc/4FTG-R6PK].

36. *See, e.g.*, *Datasets*, KAGGLE, https://www.kaggle.com/datasets [https://perma.cc/46EX-ZXMN].

37. *The bAbI Project*, FACEBOOK RESEARCH, https://research.fb.com/downloads/babi [https://perma.cc/2PS7-ZWBY].

reusing existing data is the difficulty of determining its origins, because data collected and tailored for one use may not be appropriate for another use.[38]

The utility of a dataset depends in large part on four attributes: its volume, velocity, variety, and veracity.[39] Volume indicates its size; velocity indicates its "freshness," that is, whether or not the datapoints have become outdated; variety refers to the sources of data (for example, some datasets combine information from various sources, often making them more valuable); and veracity refers to the data's accuracy.[40] As the AI is exposed to data, its system identifies patterns and can be taught to perform a huge variety of tasks, including differentiating dogs from cats,[41] bad omens from good omens,[42] or criminals who are likely to reoffend from those who are not.[43] An AI system can learn to recognize faces or emotions by sifting through huge sets of portraits from people all around the world conveying different emotions — it can distinguish between anger and sadness, or between one face and another for any number of purposes.[44] As long as there is sufficient data to train an AI system, the potential applications are endless. This process is called machine learning.[45]

---

38. Michael Mattioli, *Disclosing Big Data*, 99 MINN. L. REV. 535, 544–46 (2014).

39. Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 59 ARIZ. L. REV. 339, 345–46 (2017).

40. *Id.*

41. *Dogs vs. Cats*, KAGGLE, https://www.kaggle.com/c/dogs-vs-cats [https://perma.cc/VAP5-SWQ6].

42. Press Release, Metro Pictures, Trevor Paglen: A Study of Invisible Images, http://www.metropictures.com/exhibitions/trevor-paglen4/press-release [https://perma.cc/5AZC-JH94] ("To make the prints in *Adversarially Evolved Hallucinations*, Paglen trained an AI to recognize images associated with taxonomies such as omens and portents, monsters, and dreams. A second AI worked in tandem with the first to generate the eerie, beautiful images that speak to the exuberant promises and dark undercurrents characterizing our increasingly automated world.").

43. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/QBM7-S29B].

44. *Emotion AI Overview*, AFFECTIVA, https://www.affectiva.com/emotion-ai-overview/ [https://perma.cc/F285-8PPX].

45. For a more thorough definition of machine learning, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017) ("[M]achine learning refers to an automated process of discovering correlations (sometimes alternatively referred to as relationships or patterns) between variables in a dataset, often to make predictions or estimates of some outcome.").

Not only can AI systems "predict" outcomes that directly affect humans, but they can also create new content.[46] For example, an AI system can write novels, news articles, and, hypothetically, a court opinion or law review article.[47] An AI system can create works of art or music and produce patentable inventions.[48] Programmers pursuing writing applications for AI would train the system using datasets that teach it the structure of language, such as the interrelationships between subjects, verbs, and objects, and how particular words have congregated together previously.[49] A data trainer might accomplish this by exposing an AI system to a dataset containing written materials pertaining to a specific field, or a more generalized set of works.[50] A programmer designing an AI system that composes jazz music would expose it to a vast catalog of existing jazz recordings, which the system would break down into tiny electronic signals to learn the statistical correlations between different notes.[51] With this knowledge, the system can create new melodies that match the trajectories of preexisting ones, without "copying" earlier works.[52]

Following this initial teaching phase, the trainer must give the AI system feedback.[53] Here, the trainer will introduce new pieces of information and ask the AI system to identify its parameters.[54] The

---

46. Matthew Hutson, *How Google Is Making Music with Artificial Intelligence*, SCI. MAG. (Aug. 8, 2017, 3:40 PM), http://www.sciencemag.org/news/2017/08/how-google-making-music-artificial-intelligence [https://perma.cc/NA5B-TQHB].

47. Joe Keohane, *What News-Writing Bots Means for the Future of Journalism*, WIRED (Feb. 16, 2017, 7:00 AM), https://www.wired.com/2017/02/robots-wrote-this-story/ [https://perma.cc/TJ7Q-LZYG].

48. *See* Yanisky-Ravid & Liu, *supra* note 12, at 2224–26; Yanisky-Ravid, *supra* note 29, at 663.

49. Danny Lewis, *An AI-Written Novella Almost Won a Literary Prize*, SMITHSONIAN (Mar. 28, 2016), https://www.smithsonianmag.com/smart-news/ai-written-novella-almost-won-literary-prize-180958577/ [https://perma.cc/G9UW-YYV7].

50. *Ngram Viewer*, GOOGLE BOOKS, http://storage.googleapis.com/books/ngrams/books/datasetsv2.html [https://perma.cc/3VLQ-KSVX].

51. To listen to jazz produced by AI systems and understand how it works, see *Episode 51: AI and Intellectual Property Law, Featuring Prof. Shlomit Yanisky-Ravid*, FORDHAM INTELL. PROP. MEDIA & ENT. L.J., PODCAST (May 3, 2018) (downloaded using iTunes) [hereinafter *AI and IP Law Fordham Podcast*]; *see also* MILLION SONG DATASET, https://labrosa.ee.columbia.edu/millionsong/ [https://perma.cc/R2DN-W7XG] (2011).

52. *AI and IP Law Fordham Podcast*, *supra* note 51.

53. *See* Lehr & Ohm, *supra* note 45, at 684–88.

54. *Id.* at 685.

trainer will indicate when the system is correct and when it is wrong.[55] This feedback phase allows the AI system to hone in on its ultimate objectives, increasing the system's accuracy and efficiency.[56] Even once the system is publicly released, trainers can continue to refine the AI system by continuing to correct its errors, resulting in a feedback loop that continuously improves the AI systems' utility.[57]

Products of machine learning that depend on massive amounts of collected data promise to transform society. Such systems will not only have the capacity to predict results and meet objectives, but also be able to reform and reshape us.[58] For example, consider an AI system that presents a mix of news stories to a user, but as that user clicks on particular kinds of stories, the system alters the types of news stories it chooses to present to the user.[59] Over time, the system will learn that user's preferences and will thus reflect the news stories that the algorithm chooses to show the user.[60] Undoubtedly, the potential effects of this feedback loop are significant, as different members of society could develop dramatically different conceptions of the communal status quo. This capacity to shape opinion and perspective is precisely the reason that the workings of AI systems must be closely examined, lest there be some glitch or unseemly detail that could be used to trick and manipulate.[61] This latter concern is particularly important considering the Transparency Model discussed here. AI systems can also perpetuate disparities in employment: They can judge applicants and determine who should be offered a job

---

55. *Id.*

56. *Id.* at 696–97 (describing the "tuning process").

57. *Id.* at 699–700.

58. *See* Cohen, *supra* note 6, at 1925.

59. Julia Angwin, *On Google, a Political Mystery that's All Numbers*, WALL ST. J. (Nov. 4, 2012, 5:09 PM), https://www.wsj.com/articles/SB10001424052970203471045780991225300808 36 [https://perma.cc/2FAT-67VD] (noting an increased likelihood that a Google News user who searched for "Obama" will receive more news regarding Iran, compared to a user who searched for "Romney").

60. This mechanism is not unlike movie or clip recommendations created by algorithms for Netflix or Youtube users. *See* Libby Plummer, *This Is How Netflix's Top-Secret Recommendation System Works*, WIRED (Aug. 22, 2017), https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like [https://perma.cc/EC9K-GSR2].

61. Jihii Jolly, *How Algorithms Decide the News You See*, COLUM. JOURNALISM REV. (May 20, 2014), https://archives.cjr.org/news_literacy/algorithms_filter_bubble.php [https://perma.cc/PR78-24TM].

or which employees deserve promotions.[62] An AI system designed to predict who will be a successful employee, or who should receive a promotion, will be able to do so only after the programmer trains it using historical hiring and promoting data. But this kind of application risks perpetuating and extending historical disparities in employment, rather than ameliorating them.

In summary, AI systems can learn to perform any number of tasks typically carried out by humans. Once AI systems are exposed to massive amounts of data, they can analyze preexisting datasets, associate variables and attributes with positive and negative outcomes, and use these associations to predict, create, and decide. Through feedback, AI systems can learn from their own mistakes, improve performance, and identify more conclusive drivers of positive outcomes. But the success and reliability of AI systems depends on the kind and quality of the underlying data they are trained on. AI systems can reach undesirable social outcomes if their harmful results are deemed "positive matches" by the system because of faulty datasets, because they were not properly trained, or because of a lack oversight.

II. THE LEGAL CHALLENGES AND HURDLES OF USING BIG DATA: THE THREAT OF DISCRIMINATORY OUTCOMES AND PRIVACY VIOLATIONS

One of the main goals of this Article is to identify areas where using unevaluated datasets to train AI systems could lead to undesirable public policy outcomes. Because of the opaque nature of how AI systems function, it is often difficult (if not impossible) to determine the harms that result from using discriminatory, illegal, and unethical datasets after the fact. That being the case, this Part of the Article forms the basis for our recommendation that data providers for AI systems within the industry must take active steps to scrutinize and audit the quality of the datasets they use to train their AI systems. This kind of scrutiny early in the process is tremendously important, especially when the AI industry is mainly controlled by massive firms which are dominant in the market and have ongoing

62. For an economic study of a machine learning tool meant to predict the productivity of teachers and police officers, see Aaron Chalfin et al., *Productivity and Selection of Human Capital with Machine Learning*, 106 AM. ECON. REV. 124, 124–26 (2016).

access to massive quantities of data.[63] Moreover, where datasets are subject to buy and sell agreements, or where firms retain exclusive rights over datasets, there is a significant risk that the interests of the public will not align with the goals of the entity using the data.[64]

This Part describes two areas where a faultily-composed dataset used to train AI systems could spur adverse outcomes. Section II.A explores the ways in which partial, incomplete, wrong, or biased data can lead to discriminatory or illegal outcomes. It demonstrates that when AI is used in economic sectors where discrimination triggers liability, there is a clear need to ensure that the data is "clean." Section II.B considers how privacy requirements can pose significant obstacles to the development of AI systems and demonstrates the importance of instituting a Privacy by Design framework to ensure that privacy concerns are respected at all stages of the AI system's development, especially at the training phase.[65]

A. Discriminatory Data

1. AI and Discriminatory Data

There are many ways in which AI systems can create discriminatory outcomes by relying on "bad" data; this is becoming more and more a pressing problem, because unsurprisingly AI systems play constantly growing roles in areas of the economy where Congress prohibits discrimination.[66] AI systems, for example, conduct background checks for employment.[67] They evaluate tenants

63. MCKINSEY GLOBAL INST., *supra* note 1, at 6 (noting that companies at the "digital frontier" — online firms and digital natives such as Google and Baidu — invested an estimated $20 billion to $30 billion in AI development in 2016).

64. *See, e.g.*, *infra* notes 179–181 and accompanying text.

65. It is of note that the discussion herein, regarding adverse outcomes that result from problematic datasets, is not meant to be exhaustive — there are many other problems that arise from faulty data, such as potential copyright violations of material contained within a dataset.

66. For a discussion on the logic of anti-discrimination legislation, see generally Robert Post, *Prejudicial Appearances: The Logic of American Antidiscrimination Law*, 88 CALIF. L. REV. 1 (2000).

67. Brian Blum, *Intelligo Gives Background Checks The AI Treatment*, ISRAEL21C (July 17, 2018), https://www.israel21c.org/intelligo-gives-background-checks-the-ai-treatment/ [https://perma.cc/46N3-NVYM] ("Indeed, AI shines the brightest when it can make connections and identify patterns in the way that only a machine can . . . . Another AI advantage: it can check and recheck a target's background automatically, ensuring that any red flags on individuals or companies are discovered in real time.").

who apply for leases.[68] They determine a person's creditworthiness, and even predict future terrorist and criminal conduct.[69] In each of these examples, the objective of the AI system is to distinguish between "safe" people and those to avoid.[70] Hence, AI systems have become powerful filtering tools, sorting and categorizing persons in many areas, and their influence and dominance will surely continue to grow in significant and unpredictable ways. How AI systems accomplish these sorting tasks depends in large part on the data used by the data providers to train them. The data providers must identify for the AI system examples of good employees, tenants, or debtors.[71] Implicitly, the AI system will have to decide what personal traits make certain candidates undesirable.[72] In order to avoid running afoul of anti-discrimination laws, data providers must ensure that the variables contained within the datasets pertain specifically to the job at hand and do not contain biased, illegal, or irrelevant characteristics such as gender, race, or sexual orientation.[73]

On a more abstract level, AI developers may need to evaluate *who* gets to define the characteristics of a good employee or tenant. For example, Barocas and Selbst demonstrate the interests and stakes that policymakers must consider in the context of employment discrimination.[74] To reduce the probability that AI systems might illegally discriminate against potential employees, Barocas and Selbst have identified potential stakeholders who might be able to tackle this crucial problem, such as the employer or an auditor like the EEOC or a private third party.[75] They also identify numerous obstacles: the question of whether the employer has collected data, or

---

68. *See, e.g.*, NABORLY, https://naborly.com/ [https://perma.cc/9BAV-HAXT].

69. *See, e.g.*, ZEST FINANCE, https://www.zestfinance.com/ [https://perma.cc/KR7D-5NDQ]; *see also supra* notes 2 and 3.

70. *See generally* Desai & Kroll, *supra* note 28.

71. *See* Lehr & Ohm, *supra* note 45, at 672–73.

72. *Id.* at 665 ("If programmers specify output variables in ways that make members of certain demographic groups more likely than others to have 'advantageous' outcomes, discrimination can be introduced.").

73. It is worth mentioning that anti-discrimination statutes do seem to imply that AI operators should choose unbiased datasets that do not result in unequal, prejudicial outcomes. *See* Barocas & Selbst, *supra* note 19, at 694–714. However, legal obstacles, such as trade secret protections, make it very difficult to actually achieve the goals of non-discriminatory data. *See generally* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018) (discussing how trade secrets prevent scrutinizing sentencing algorithms and other uses of AI in the criminal justice system).

74. *See* Barocas & Selbst, *supra* note 19, at 694–714.

75. *Id.* at 718–19.

if it resorted to data collected by third parties; whether the model used is internally developed, or if it was purchased from a third party.[76] They note that third parties that specialize in human resources may prove more than able to audit the employment decisions made by AI models.[77]

2. Disparate Impact

Numerous federal statutes prohibit discrimination.[78] There are two main types of unlawful discrimination: intentional discrimination and disparate impact.[79] Disparate impact claims arise where a policy, although facially neutral, creates a statistical imbalance that adversely impacts a protected group:[80]

> Disparate impact is not concerned with the intent or motive for a policy; where it applies the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.[81]

Under a theory of disparate impact, a plaintiff need not demonstrate any animus on the part of the defendant.[82] Instead, all the standard requires is that a protected group is disproportionately affected by a practice or policy relative to other groups.[83] There are generally four elements to a disparate impact claim: (1) statistical imbalances that indicate an adverse impact on a protected group, caused by (2) a facially neutral policy, (3) which was "artificial, arbitrary, and unnecessary," supported by (4) factual allegations that

---

76. *Id.*

77. *See id.*

78. *See* Shlomit Yanisky-Ravid, *Gender Pay Gap: Blaming the Victims — Are Women Responsible?!*, *in* NYU ANNUAL CONFERENCE ON LABOR 91, 96–103 (67th ed. 2015) (reviewing antidiscrimination legislations regarding gender and the gender pay gap) (on file with author).

79. *See* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507, 2513 (2015) (noting the distinction between disparate treatment and disparate impact).

80. Barocas & Selbst, *supra* note 19, at 694.

81. *Id.*

82. *See id.* at 701.

83. *See id.*; *see also Inclusive Communities*, 135 S. Ct. at 2518 ("[A]ntidiscrimination laws must be construed to encompass disparate-impact claims when their text refers to the consequences of actions and not just to the mindset of actors, and where that interpretation is consistent with statutory purpose.").

indicate a "robust causation" between the policy and the disparity.[84] When these elements are met, it is possible to demonstrate discrimination through statistical analysis by showing that a facially neutral policy resulted in a disparate impact.[85]

The U.S. government regulates discrimination by economic sector and heightens prohibitions in specific contexts such as employment, housing, and consumer finance.[86] The Fair Housing Act of 1968 (FHA), for example, prohibits discrimination in the buying or selling of homes on the basis of "race, color, religion, sex, familial status, or national origin."[87] The Age Discrimination in Employment Act of 1967 (ADEA) prohibits discrimination against people over the age of forty in employment decisions.[88] There are many other anti-discrimination statutes in various areas of American jurisprudence that read very similarly to the FHA or ADEA.[89] The overarching ideology behind these statutes is that the characteristics of a person belonging to a protected group ought to be irrelevant to the selection, evaluation, or compensation of that person; by making those factors irrelevant, the law can eliminate "stubborn but irrational prejudice."[90] It is worth noting that the statutes mentioned above prohibit both intentional discrimination and practices that produce disparate impacts.[91]

While there has not been extensive litigation regarding AI systems and discrimination, existing cases make clear that decision-makers are

---

84. *See, e.g.*, *Inclusive Communities*, 135 S. Ct. at 2522–24; *see also* Robert G. Schwemm, *Fair Housing Litigation After* Inclusive Communities*: What's New and What's Not*, 115 COLUM. L. REV. SIDEBAR 106 (2015), https://columbialawreview.org/content/fair-housing-litigation-after-inclusive-communities-whats-new-and-whats-not/ [https://perma.cc/9GD6-H7VX].

85. *See generally* Yanisky-Ravid, *supra* note 78.

86. *See infra* notes 87–90.

87. *See* 42 U.S.C. § 3604 (1988).

88. *See* 29 U.S.C. §§ 623, 631 (2016); *see also* Smith v. City of Jackson, Miss., 544 U.S. 228, 232 (2005) (noting that Congress enacted the ADEA in response to a report from the Department of Labor, which stated that arbitrary discrimination on the basis of age was occurring).

89. *See* George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2318 (2006); Christine Jolls, *Antidiscrimination & Accommodation*, 115 HARV. L. REV. 642, 643 (2001) ("The canonical idea of 'antidiscrimination' in the United States condemns the differential treatment of otherwise similarly situated individuals on the basis of race, sex, national origin, or other protected characteristic.") (internal quotations omitted).

90. Lam v. Univ. of Haw., 40 F.3d 1551, 1563 (9th Cir. 1994); *see also* Post, *supra* note 66, at 10.

91. *See* Rutherglen, *supra* note 89, at 2330–31 (2006). *See generally* Schwemm, *supra* note 84.

prohibited by law from using a variable that is irrelevant to job performance and produces disparate impacts. When the Southeastern Pennsylvania Transit Authority's police department instituted minimum running speed requirements, ostensibly to increase the quality of the police force, a district court struck those requirements down on the basis of their disparate impact on women applicants.[92] A fire department was prohibited from considering personal contacts and familial relationships when evaluating candidates because of the negative disparate effect this would have on black candidates.[93] These kinds of prohibitions are reminiscent of AI systems that evaluate social media accounts for tenants or run employment background checks.[94] Existing precedent logically extends to prohibit utilizing variables that have a disparate impact on protected groups in creating, developing and relying upon machine learning systems as decision-makers.

These cases demonstrate that biased results can occur despite any explicit reference to a protected group, and this should concern AI system operators who work in regulated sectors. ZestFinance, a lending company that has an AI system to determine whether someone should receive a loan, serves as a good example of how the cases described above can apply to AI when the system relies on new datapoints.[95] Its system is trained to flag when potential clients quickly scroll through the terms and conditions and interprets someone who divulges their social media connections as a riskier applicant than someone who does not.[96] These attributes may or may not be good indicators of creditworthiness — scrolling through terms and conditions, and the speed at which you do so, could be a function of any number of variables, such as device type, education level, or financial desperation.

---

92. Jolls, *supra* note 89, at 656–57 (discussing Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478 (3d Cir. 1999)).

93. *Id.* at 657 (discussing Banks v. City of Albany, 953 F. Supp. 28, 33–36 (N.D.N.Y 1997)).

94. *See generally* Blum, *supra* note 67. Consider Fama, for example, a company that purports to be able to utilize AI to "identify threats related to sexual harassment, discrimination, theft of sensitive information, or other types of people risk that could destabilize your organization," or to provide "role specific screening packages [that] can help you identify those individuals likely to excel in the organization." *Product Overview*, FAMA, https://www.fama.io/product-overview/ [https://perma.cc/9MFF-FF63].

95. *See* Hurley & Adebayo, *supra* note 21, at 164–65.

96. *See id.*

3. The Impact of Bad Data: Biased, Partial, or Wrong

When the training data is unrepresentative of the environment it will operate in, discriminatory outcomes can occur. This can come about through a host of issues. The dataset itself may be over-inclusive or under-inclusive of certain segments of the population, creating sample-size issues that produce disparate impacts.[97] The dataset could contain preexisting biases that lead the AI system to inherit society's biases, habits, and beliefs, and perpetuate already persisting discrimination.[98] For example, consider an AI system meant to predict the likelihood of an employee's future success by learning from the dataset of previous achievements. The dataset of those previous achievements "might be tilted against minorities given their lack of past success because of a harsh work environment resulting from discriminatory attitudes, or the lack of minority hiring altogether."[99] Not only does this AI system rely on faulty data, the context in which the data was collected and the present circumstances may have changed dramatically, reducing the predictive worth of the dataset.[100] "Confounding covariates," such as the use of ZIP codes to predict the worthiness of applicants, can drive an AI system to produce racist outcomes, even though the system was never taught what race is.[101] Also known as proxies or "redundant encodings," confounding covariates can render significant efforts to counteract discrimination wholly ineffective, and ultimately produce disparate impacts.[102] In other words, relying on confounding covariates raises the possibility that AI systems could resort to making decisions based on a datapoint that disproportionately harms one group over another, even if the AI system was never taught concepts of race, gender, sexuality, age, or other protected characteristics.

Bad data can also lead to faulty decision-making processes and conclusions.[103] In particular, missing and inaccurate data can

---

97. *See id.* at 171.

98. *See* Batya Friedman & Helen Nissenbuam, *Bias in Computer Systems*, 14 ACM TRANSACTIONS ON INFO. SYSTEMS 330 (1996).

99. Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375, 1392 (2014).

100. *See* Lehr & Ohm, *supra* note 45, at 684–85.

101. *See* Zarsky, *supra* note 99, at 1394–95.

102. *See* Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1036–37 (2017).

103. *See generally* James T. Graves et al., *Big Data and Bad Data: On the Sensitivity of Security Policy to Imperfect Information*, 83 U. CHI. L. REV. 117 (2016) (discussing the myriad ways that bad data leads to bad decisions in the security context).

potentially lead an AI system to draw invalid inferences.[104] For example, consider a machine learning algorithm that evaluates applications to higher education institutions. Historically underrepresented minorities at colleges and universities may be denied admission simply because of their preexisting lack of representation in the data sample. The same can be said for non-traditional applicants with uncommon life trajectories. These kinds of historically underrepresented applicants may not be evaluated as holistically by an AI system as they would be by a human admissions officer. In the best-case scenario, over time those errors would be corrected by the machine learning algorithm's feedback loop. But even in that situation, there would nonetheless remain a potentially lengthy interim period where invalid inferences are drawn from lacking data inputs. By requiring real transparency of the sources of the data, such problems could be reduced or avoided.[105] Under the Transparency Model proposed here, the data provider has to critically scrutinize the data they use to train the AI system and ensure that it will not continue to perpetuate historically discriminatory imbalances.

4. Discrimination in the Feedback

AI systems can be taught to reach discriminatory or biased conclusions not only from faulty underlying datasets, but also during the feedback phase. This can occur in two ways: through a transmission of the trainer's biases, or through a failure on the part of the trainer to correct the AI's mistakes. The former of these is not the consequence of the AI system or the underlying data as much as it is on the shoulders of discriminatory human conduct. However nefarious this conduct is, it is not as relevant to the scope of this work. In the latter case, the AI system learns using a dataset that is not per se discriminatory — it is the process used to teach the system that causes it to produce biased results. The trainer may adjust the algorithm to match the dataset in such a way that it "overfits" the dataset, forcing the AI system to create arbitrary connections that result from the randomness of the feedback it received.[106] In such a case, the AI system *creates* rather than *identifies* patterns in ways that do not recognize proper cause and effect, inferring false interrelationships between variables that are, in fact, not related at all.

---

104. *See id.* at 119–20.
105. *See* Zarsky, *supra* note 99, at 1395.
106. *See* Lehr & Ohm, *supra* note 45, at 714.

5. Unmoored and Independent AI Systems that Autonomously Seek Data

The feature that most distinguishes AI systems from traditional algorithms is their independence: AI systems are capable of searching for new relevant datasets from areas such as social networks, internet sites, blogs, and other data that exists online.[107] Therefore, once in operation, an AI system can become unleashed from and move beyond its original training data as it collects new information. An unrestrained AI system can create what is known as "arbitrariness by algorithm" — it might predict outcomes based on inferences, correlations, and groupings of different individuals together according to data found online, which might be filled with arbitrary and misleading datapoints.[108] Consider, for example, a data provider that initially exposes and trains an AI system to hire or promote employees without considering gender or race. As the AI system continues to evolve on its own, receiving employee data regarding hiring and promotions, it could begin to identify other factors such as whether employees play football or dance ballet. Because these other characteristics are often tied to race, gender, sexuality, class, and the like, these unrelated variables might lead the system to become biased on the basis of these characteristics, which will ultimately result in biased outcomes. Even though anti-discrimination statutes clearly prohibit AI data providers and trainers from teaching an AI system to purposely discriminate, that the system will autonomously become prejudicial is a very real possibility — one that falls within the parameters of anti-discrimination legislation because of the disparate impact that the AI system would ultimately produce.[109]

6. From Theory to Practice: The Inability of Existing Laws to Control AI Systems

While it is theoretically possible that the disparate impact doctrine could apply broadly to improperly assembled datasets, there is reason to conclude that it may not. Perhaps the AI system operator could escape liability by claiming a business necessity, a defense to disparate impact that arises where the practice at issue is the only way of

---

107. *See* Yanisky-Ravid & Liu, *supra* note 12, at 2223–29.

108. *See* Zarsky, *supra* note 99, at 1408–09.

109. *See* Friedman & Nissenbuam, *supra* note 98, at 330, 335 (discussing "Emergent Bias" that arises from the use of the computer system, rather than being preexisting or technical).

accomplishing a worthwhile objective.[110] In the case of an AI operator, she might demonstrate that the data used to train the AI system contained necessary variables related to the AI's objective, thereby potentially defeating a disparate impact claim.[111] Barocas and Selbst argue, for example, that the collection of vast amounts of data to use in the employment context is not effectively counteracted by the requirements of anti-discrimination legislation in the employment sector.[112] Additionally, the disparate impact doctrine cannot address the myriad issues that are likely to arise from faulty data in AI — the doctrine has been relegated to specific practices such as employment, housing, and lending practices.[113] Besides, many practices can produce outcomes that may not rise to prohibited discrimination but nonetheless offend society. For example, one study demonstrated that, when a user's "ad preference settings were set to female, a user saw 'fewer instances of an ad related to high-paying jobs than [when preferences were set] to male,'" demonstrating that somewhere in the system, the AI made a biased connection between higher earnings and male job applicants.[114]

It is vital to recognize that even though these problems are detrimental to society and risk increasing inequality, it is possible to train AI systems to reduce the potential for bias.[115] Rather than perpetuating existing biases and disparities, an AI system could learn to avoid such outcomes by being taught to identify where bias or discrimination may be driving particular decisions.[116] One AI system could even audit another to "ensure that [data] is used only for its intended purpose," and to "identify algorithmic outcomes that are unfair and discriminatory."[117]

---

110. *See* Smith v. City of Jackson, Miss., 544 U.S. 228, 244 (2005) (explaining the business necessity defense).

111. *See* Lehr & Ohm, *supra* note 45, at 666.

112. *See generally* Barocas & Selbst, *supra* note 19.

113. *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 704–05 (2006) (arguing that the disparate impact doctrine has had limited impact).

114. Desai & Kroll, *supra* note 28, at 17–18.

115. Lehr & Ohm, *supra* note 45, at 704–05.

116. Philip Hacker & Bilyana Petkova, *Reining in the Big Promise of Big Data: Transparency, Inequality, and New Regulatory Frontiers*, 15 NW. J. TECH. & INTELL. PROP. 1, 13 (2017) (describing anecdotal evidence of Amazon charging higher prices for Mac users than Windows users, suggesting that because the average Mac user is wealthier than a PC user, income inequality is reduced).

117. Andrea Scripa Els, Note, *Artificial Intelligence as a Digital Privacy Protector*, 31 HARV. J.L. & TECH. 217, 224 (2017).

7. Examples and Consequences of Discriminatory Behavior by AI Systems

In 2016, Microsoft released Tay, its new social media AI system, with great fanfare.[118] Less than a day later, Microsoft apologized and removed Tay from Twitter.[119] During that single day, Tay had learned how to be racist, genocidal, and a white supremacist.[120] As it turns out, "a small subset" of Twitter users "exploited a vulnerability" in the program,[121] apparently a problematic "repeat after me" function.[122] Though this may seem like an isolated incident, it is hardly the only instance of misbehavior.[123] It also highlights the broader risks regarding artificial intelligence and discrimination.[124]

Furthermore, while Tay is an obvious example of how AI applications can malfunction if given the wrong training data (i.e. speech by online trolls[125]), in other instances it is much more difficult to trace whether the undesired result came from a discriminatory dataset or a prejudicial trainer. For example, consider an algorithm

---

118. Abby Ohlheiser, *Trolls Turned Tay, Microsoft's Fun Millennial AI Bot, Into a Genocidal Maniac*, WASH. POST (Mar. 25, 2016), https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/ [https://perma.cc/K6XZ-SVY5].

119. *Id.*

120. *Id.*

121. Peter Lee, *Learning from Tay's Introduction*, OFFICIAL MICROSOFT BLOG (Mar. 25, 2016), https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ [https://perma.cc/CT9A-VYHJ].

122. Ohlheiser, *supra* note 118.

123. *See* Jana Kasperkevic, *Google Says Sorry for Racist Auto-Tag in Photo App*, GUARDIAN (July 1, 2015), https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app [https://perma.cc/SUY9-XJC3]; Matt Day, *How LinkedIn's Search Engine May Reflect a Gender Bias*, SEATTLE TIMES (Aug. 31, 2016), https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/ [https://perma.cc/F3U2-HVZZ].

124. *See* Stephen Buranyi, *Rise of the Racist Robots — How AI Is Learning All Our Worst Impulses*, GUARDIAN (Aug. 8, 2017), https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses [https://perma.cc/D3VK-ND5G]; Christina Couch, *Ghosts in the Machine*, PBS (Oct. 25, 2017), http://www.pbs.org/wgbh/nova/next/tech/ai-bias/ [https://perma.cc/WUP9-JBRD]; Sophia Chen, *AI Research Is in Desperate Need of an Ethical Watchdog*, WIRED (Sept. 18, 2017), https://www.wired.com/story/ai-research-is-in-desperate-need-of-an-ethical-watchdog/ [https://perma.cc/T6BS-MYYU].

125. A troll is someone who "antagonize[s] (others) online by deliberately posting inflammatory, irrelevant, or offensive comments or other disruptive content." *Troll*, MERRIAM-WEBSTER (Online ed. 2019), https://www.merriam-webster.com/dictionary/troll [https://perma.cc/2L8Q-5948].

that decides SAT prep courses should cost twice as much in Asian-majority communities than others — is it doing so because those communities are Asian, or for some other reason?[126] What if a disparity occurs because of the underlying sample size, such as facial recognition applications that identify white male faces much more effectively than women and other races?[127]

Discriminatory outcomes pose both legal and reputational dangers to AI systems operators. Where statutes prohibit business practices that cause a disparate impact for protected groups without a valid reason, such as a business necessity defense, an AI operator should ensure that its systems are run in compliance with those statutes. Additionally, the specter of discrimination can also lead to serious reputational damage, which should incentivize AI operators to consider possible vulnerabilities in their AI training process that could produce discriminatory outcomes.

The above sections demonstrate how exposing an AI system to biased, incomplete, or wrong data can result in discrimination. As AI systems become cheaper to implement and more widely used, these risks will increase. Technology firms should consider these risks when developing AI systems and avoid inappropriate uses of data that may violate anti-discrimination laws or result in an AI malfunction that could damage their reputation. The Transparency Model would help increase public trust by educating users, and incentivize the industry to scrutinize and verify the data its AI systems depend on.

There is one significant barrier to understanding how and when it is more likely for an AI system to produce discriminatory results: lack of transparency. Currently, consumers are unable to know with certainty whether a given negative outcome resulted from an AI system built on biased data, or due to some other reason. The next section examines how we can resolve this issue by grounding the public's right to an explanation in a right to privacy. Privacy rights function both as a means to understanding whether discrimination is occurring, and as an end in themselves so that individuals can confidently ascertain whether private information about them is being misused in ways that they did not foresee. By encouraging

---

126. Julia Angwin & Jeff Larson, *The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review*, PROPUBLICA (Sept. 1, 2015), https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review [https://perma.cc/ZGY6-SE3G].

127. *See* Couch, *supra* note 124 (discussing the popular benchmark Labeled Faces in the Wild used by many well-known tech firms to measure algorithm performance for facial recognition and the lack of diversity in the sample).

transparency and scrutiny of AI systems, the Transparency Model can help consumers hold operators accountable, such that they will better prevent discriminatory results and avoid privacy violations.

B. Data and Privacy: Invasive and Pervasive Data

The ways that AI trainers use data pertaining to individuals invariably raises privacy concerns. Individuals are not always aware of the sheer magnitude of data that others possess about them, and therefore likely are not aware of how this data is used by AI systems. Privacy plays a pivotal role in the Transparency Model for two interlocking reasons. For one, transparency would incentivize the AI industry to avoid violating societal privacy expectations. Conjointly, transparency can empower individuals to better control their data, and will enable them to advocate for their data privacy more effectively.

AI systems cannot reliably function without the huge amounts of data used to train them. At present, there is no shortage of data that relates to the most personal details about almost all people.[128] Some companies, such as Acxiom, make data available to individual users without a special request; others, such as Facebook, require that consumers send requests specifically to Facebook before it will release the data it has collected on them.[129] Consumers are often unaware of the detailed information that AI system operators possess, which is drawn from multiple sources. This creates two layers of secrecy: Not only do consumers not know what data exists, but they also do not know how AI systems use it. Thus, consumers cannot opt out of being "judged" by a computer, nor are they able to control the information that others hold about them. This section explores the role that privacy can play in helping policymakers craft a response to the proliferation of AI in all aspects of our society. Following this analysis of privacy, Part III discusses the Transparency Model in detail, which will help educate consumers and incentivize AI system operators (and others) to consider what types of data they

---

128. Natasha Singer, *What You Don't Know About How Facebook Uses Your Data*, N.Y. TIMES (Apr. 11, 2018), https://www.nytimes.com/2018/04/11/technology/facebook-privacy-hearings.html [https://perma.cc/D48E-UETK].

129. *Compare* ABOUT THE DATA BY ACXIOM, https://aboutthedata.com/portal [https://perma.cc/EBA6-BLM9], *with* Louise Matsakis, *What to Look for in Your Facebook Data — And How to Find It*, WIRED (Mar. 28, 2018), https://www.wired.com/story/download-facebook-data-how-to-read/ [https://perma.cc/HL3S-3TJY].

should use in training AI, and, on a more fundamental level, how much of that data they should use.

Many scholars have recognized that privacy is difficult to define as a legal concept.[130] For purposes of this Article, the notion of privacy focuses on the psychological aspects of the term and, more specifically, the perceived personal "balloon," or a private sphere, which centers on autonomy, freedom, and the creation of relationships.[131] Accordingly, this Article envisions privacy in a unique way by relying on the conceptual balloon of privacy that surrounds each and every one of us.[132]

Efforts to regulate data collection and to protect privacy have recently become more commonplace and comprehensive. The European Union's General Data Protection Regulation (GDPR), which took effect on May 25, 2018, calls for protecting privacy rights in the collection, use, and maintenance of individual data and imposes these same mandatory obligations on U.S. firms that have commercial relationships with European firms or citizens.[133] The U.S. Federal Trade Commission (FTC) and the Office of the Attorney General of California have recognized the importance of privacy and the need to design policies that protect privacy as part of the data industry's inner processes of production, known as Privacy by Design.[134]

---

130. *See, e.g.*, Richard A. Posner, *The Right of Privacy*, 12 GA. L. REV. 393, 393 (1977) ("The concept of 'privacy' is elusive and ill defined. Much ink has been spilled in trying to clarify its meaning."). *But see* James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151, 1153–55 (2004) (arguing that the fact that privacy is embarrassingly difficult to define and differs from one culture to another undermines the importance of keeping privacy as a tool to protect personhood). According to Daniel Solove, the multitude of definitions produced by scholars can be narrowed down to six basic conceptualizations of privacy: (1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control of personal information; (5) personhood; and (6) intimacy. Daniel J. Solove, *Conceptualizing Privacy*, 90 CALIF. L. REV. 1087, 1094 (2002).

131. *See* James Rachels, *Why Privacy Is Important*, 4 PHIL. & PUB. AFF. 323, 326–31 (1975). *See generally* Shlomit Yanisky-Ravid, *To Read or Not to Read: Privacy Within Social Networks, the Entitlement of Employees to a Virtual Private Zone, and the Balloon Theory*, 64 AM. U. L. REV. 53, 80–86 (2014) (introducing the balloon theory).

132. Yanisky-Ravid, *supra* note 131, at 83–84.

133. *See generally* Council Regulation 2016/679, of the European Parliament and of the Council of 27 Apr. 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, 2018 O.J. (L 119), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 [https://perma.cc/AE85-XBKD].

134. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 22–23 (Mar. 2012), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-

Privacy rights are also well-recognized in international law. Article 12 of the Universal Declaration of Human Rights states: "No one shall be subjected to arbitrary interference with his privacy, family, home, or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks."[135] This text is memorialized in Article 17 of the International Convention On Civil and Political Rights of 1966.[136] Similarly, Article 8(1) of the European Convention on Human Rights ensures that "[e]veryone has the right to respect for his private and family life, his home and his correspondence," and allows for interference with this heralded right to privacy in extremely limited circumstances.[137]

One does not need to provide a precise definition of privacy or determine whether a right to privacy exists under U.S. law in order to advocate that American consumers are owed proper privacy protections. For one, a reasonable expectation of privacy is justified through psychological approaches, such as the aforementioned balloon of privacy or the so-called Magnet Field Theory, which suggests that there is a sphere of privacy that always surrounds humans, even in cyberspace arenas such as social networks.[138]

---

commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf [https://perma.cc/DQF5-AG6A]; CAL. DEP'T OF JUSTICE, OFFICE OF THE ATTORNEY GEN., PRIVACY ON THE GO: RECOMMENDATIONS FOR THE MOBILE ECOSYSTEM 4 (Jan. 2013), https://oag.ca.gov/sites/all/files/agweb/pdfs/privacy/privacy_on_the_go.pdf [https://perma.cc/ZPC7-Z29E] ("Our recommendations, which in many places offer greater protection than afforded by existing law, are intended to encourage all players in the mobile marketplace to consider privacy implications at the outset of the design process.").

135. G.A. Res. 217 (III) A, Universal Declaration of Human Rights (Dec. 10, 1948), http://www.un.org/en/universal-declaration-human-rights/index.html [https://perma.cc/N2DM-YPRF].

136. International Convention on Civil and Political Rights art. 17, Dec. 19, 1966, 999 U.N.T.S. 171, http://ec.europa.eu/justice/policies/privacy/docs/16-12-1996_en.pdf [https://perma.cc/WQ56-26T7].

137. Convention for the Protection of Human Rights and Fundamental Freedoms art. 8, Nov. 4, 1950, 213 U.N.T.S. 222, http://www.echr.coe.int/Documents/Convention_ENG.pdf [https://perma.cc/4BBE-EDJ6]. For a discussion of the meaning of private life under the treaty, see IVANA ROAGNA, COUNCIL OF EUROPE HUMAN RIGHTS HANDBOOKS: PROTECTING THE RIGHT TO RESPECT FOR PRIVATE AND FAMILY LIFE UNDER THE EUROPEAN CONVENTION OF HUMAN RIGHTS 12 (2012), https://www.echr.coe.int/LibraryDocs/Roagna2012_EN.pdf [https://perma.cc/P7JF-SCK5].

138. Yanisky-Ravid, *supra* note 131 (describing the psychological importance of having a private sphere surrounding individuals in Cyberspace — as the Balloon Theory or the Magnet Field Theory — and claiming that privacy is justified with

Additionally, a reasonable right to privacy can be justified for efficiency reasons.[139] Data breaches and the release or compromise of data induce anxiety in the subjects of that compromised data.[140] Under U.S. law, communal expectations of privacy can stem from statutes.[141] However, individuals also have expectations to privacy that are normative in nature. [142] These stem from what people choose to divulge to others, and what they anticipate will happen to that information. Those normative expectations can erode over time — as AI systems become more complex and more capable of drawing upon multiple sources of data to paint new pictures of who someone is as a person, it is worth exploring whether, and to what extent, AI systems violate both statutory requirements and normative expectations of privacy.[143] Proper privacy protections depend primarily on two types of procedures: gaining informed consent for data collection, and the anonymization of collected data.[144]

When AI systems use massive amounts of data drawn from multiple sources, the potential for privacy violations increases. In certain areas of the economy, heightened *data* protections apply because policymakers decided that those specific fields warrant heightened *privacy* protections, which surmount other considerations such as the availability of cheaper products, or more efficient

---

respect to employees particularly for efficiency reasons, such as maximizing productivity and creating positive incentives).

139. *Id.*; *see also* Shlomit Yanisky-Ravid & Ben Zion Lahav, *Public Interest vs. Private Lives — Affording Public Figures Privacy in the Digital Era: The Three Principles Filtering Model*, 19 U. PA. J. CONST. L. 975, 978–79 (2017) (discussing the importance of privacy, especially in the digital era, in regards to public figures); Shlomit Yanisky Ravid & Amy Mittelman, *Gender Biases in Cyberspace: A Two-Stage Model, the New Arena of Wikipedia and Other Websites*, 26 FORDHAM INTELL. PROP. & ENT. L.J. 381, 401–12 (2015) (examining the "price" women pay when being discriminated against, and the harm suffered from harassment in the virtual spheres, such as in cases of revenge porn).

140. Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 TEX. L. REV. 737, 741 (2018) (stating that most courts dismiss data breach lawsuits because they fail to allege harm, and that this difficulty largely stems from the fact that data breach harms are intangible, risk-oriented, and diffuse for a court to assess risk and anxiety in a concrete and coherent way).

141. Colin Shaff, *Is the Court Allergic to* Katz*? Problems Posed by New Methods of Electronic Surveillance to the "Reasonable-Expectation-of-Privacy" Test*, 23 S. CAL. INTERDISC. L.J. 409, 438–39 (2014).

142. *See* Yanisky-Ravid, *supra* note 131, at 84.

143. *Id.* at 99–100.

144. Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Procedural Privacy Protections*, 57 COMM. ACM 31, 31 (2014) ("Privacy protections for the past 40 years have concentrated on two types of procedural mitigation: informed consent and anonymization.").

services.[145] In order to protect against violations of societal expectations of privacy, AI operators should make transparent the sorts of data they use to train AI systems, where the data was collected, how it is used, and how it is protected. Further, AI operators ought to take steps to ensure that such private data is not misused in ways that can trample privacy expectations or produce unfair outcomes.

1. AI and U.S. Privacy "Islands": Healthcare, Finance, and Children

This section discusses privacy in the context of "islands" of well-protected rights under U.S. law. First, we describe the advances to which AI systems have contributed in the healthcare and pharmaceutical industry and highlight the importance of data privacy in that arena. Second, we discuss how statutes require privacy with respect to children and education and note that those requirements reduce the ability of AI systems to increase in prominence in that sector. Third, we examine statutory requirements of privacy and disclosure in the Fair Credit Reporting Act (FCRA). Fourth, we identify general normative expectations of privacy and advocate for various manifestations of those expectations that AI trainers should consider when deciding what data to use in training AI systems.

a. AI in the Healthcare Field

The medical sector is bursting with developing AI systems.[146] Applications such as diabetes monitoring,[147] medical image

145. *See* Yanisky-Ravid, *supra* note 131, at 99–101.

146. Daniel Faggella, *Machine Learning Healthcare Applications — 2018 and Beyond*, TECHEMERGENCE, https://www.techemergence.com/machine-learning-healthcare-applications/ [https://perma.cc/L8CS-W7AD]; *see also* Charles Ornstein & Katie Thomas, *Sloan Kettering's Cozy Deal with Start-Up Ignites a New Uproar*, N.Y. TIMES (Sept. 20, 2018), https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html?rref=collection%2Fsectioncollection%2Fhealth&action=click&contentCollection=health&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=sectionfront [https://perma.cc/VQ4K-3SYJ] (discussing an artificial intelligence start-up founded by three insiders at Memorial Sloan Kettering Cancer Center, with $25 million in venture capital and the promise that it might one day transform how cancer is diagnosed. The company, Paige.AI, is one of many that apply artificial intelligence to healthcare).

147. *Medtronic and IBM Watson Health Partner to Develop New Ways to Tackle Diabetes*, MEDTRONIC, http://www.medtronic.com/us-en/about/news/ibm-diabetes.html [https://perma.cc/FP4S-B6TR].

analysis,[148] diagnosis of cancer and other illnesses,[149] and individualized medicinal regimens[150] are just the tip of the iceberg with respect to possible medical applications of AI. One consultant estimates that AI systems could generate up to a hundred-billion dollars in annual value in the U.S. healthcare system alone.[151] A McKinsey Global Institute publication stated that AI in healthcare could lead to quicker diagnoses, better treatment plans, and improved health insurance.[152] One significant driver of AI development in this field has been the ability to collect more and more medical data.[153] One potential issue is whether the data collected is recorded in standardized formats — presently, much data is recorded idiosyncratically, according to the needs of the collector, limiting its potential use.[154] Regulation of medical devices poses significant challenges to developers seeking to train AI systems with compliant

148. *Project InnerEye — Medical Imaging AI to Empower Clinicians*, MICROSOFT, https://www.microsoft.com/en-us/research/project/medical-image-analysis/ [https://perma.cc/CKP4-HM6B]; *see also Watson Oncology*, MEMORIAL SLOAN KETTERING CANCER CTR., https://www.mskcc.org/about/innovative-collaborations/watson-oncology [https://perma.cc/D7SW-M8KG].

149. Cooper Katz, *AI Created to Transform Cancer Diagnosis and Treatment by Applying Artificial Intelligence to Pathology*, BUS. WIRE (Feb. 5, 2018), https://www.businesswire.com/news/home/20180205005557/en/Paige.AI-Created-Transform-Cancer-Diagnosis-Treatment-Applying [https://perma.cc/H49B-GWJX] ("Currently, highly-trained pathologists interpret numerous individual glass slides using a microscope to analyze tissues and cells. For instance, there can be as many as 60 slides in one breast biopsy, though often only a few are relevant to the diagnosis. The manual nature of this work is inherently inefficient and can contribute to diagnostic variability. Technology to convert glass slides to digital images has existed for over a decade, but it has not been broadly adopted because digitization alone does not allow pathologists to improve workflow." AI technology could provide the missing link to help pathologists make decisions with greater speed, accuracy, objectivity and reproducibility — all at a lower cost.); *see also* Ornstein & Thomas, *supra* note 146 (arguing that the goal of using AI tehnology is to provide predictive data and help cancer physicians around the country).

150. Jennifer Kite-Powell, *Artificial Intelligence and Data Driven Medicine*, FORBES (July 27, 2016), https://www.forbes.com/sites/jenniferhicks/2016/07/27/artificial-intelligence-and-data-driven-medicine/#740a31c73069 [https://perma.cc/U689-5M5Y].

151. Jamie Cattell et al., *How Big Data Can Revolutionize Pharmaceutical R&D*, MCKINSEY (Apr. 2013), https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d [https://perma.cc/977P-4ZPF].

152. *See* MCKINSEY GLOBAL INST., *supra* note 1, at 63.

153. *ResearchKit & CareKit*, APPLE, https://www.apple.com/researchkit/ [https://perma.cc/63E5-ZNU2].

154. MCKINSEY GLOBAL INST., *supra* note 1, at 64.

data,[155] so much so that AI systems have been created to reduce regulatory risk with medical data.[156] This is still a nascent field, and its effectiveness is not set in stone. Watson, IBM's AI infrastructure, for example, has recommended incorrect cancer treatments at times and is often disagreed with by doctors around the world.[157]

Congress passed the Healthcare Insurance Portability and Accessibility Act of 1996 (HIPAA) to address the issue of data privacy in the healthcare sector.[158] There are two primary aspects of HIPAA: the Privacy Rule, which governs the content of data, and the Security Rule, which concerns how an entity protects data.[159] The Privacy Rule regulates individually identifiable health information collected by healthcare providers.[160] It lists situations where a healthcare provider may use or disclose health information, and requires protocols and procedures that protect health information from unauthorized access, use, or disclosure.[161] The Privacy Rule is designed to balance individual privacy rights against the societal goals of "oversight, research, law enforcement, public health and safety."[162] The Security Rule only applies to protected health information in electronic format, and has three central tenets — that the data is kept confidential, available to authorized persons only, and that the integrity of the data is assured.[163] Unlike the Privacy Rule, "the Security Rule created the standards for ensuring that only those who should have access to [electronic protected health information] will in fact have access."[164]

---

155. *See* Jonathan Kay, *How Do You Regulate a Self-Improving Algorithm?*, ATLANTIC (Oct. 25, 2017), https://www.theatlantic.com/technology/archive/2017/10/algorithms-future-of-health-care/543825/ [https://perma.cc/W5FW-85P9].

156. *See* APTIBLE, https://www.aptible.com/ [https://perma.cc/GNW3-YX7A].

157. *See* Carly Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ [https://perma.cc/5JH4-TJ6N].

158. *See generally* C. STEPHEN REDHEAD, CONG. RESEARCH SERV., R43991, HIPAA PRIVACY, SECURITY, ENFORCEMENT, AND BREACH NOTIFICATION STANDARDS (2015).

159. Health Insurance Reform: Security Standards, 68 Fed. Reg. 8334 (Feb. 20, 2003) (adopting "standards for the security of electronic protected health information").

160. 45 C.F.R. § 164.500 *et seq.* (2019).

161. *Id.* § 164.506.

162. REDHEAD, *supra* note 158, at 3.

163. *Id.* at 11.

164. *Id.*

If an entity decides to use or disclose protected health information, it must do so only to the extent necessary, though there are numerous exceptions to that "minimum necessity" standard.[165] Generally, the minimum necessity determination is conducted at a policy and procedure level, rather than for each distinct use or disclosure of protected health information.[166] For example, a healthcare organization would identify persons and departments that require protected information and limit their access to the nature of the information needed. Where the protected health information is electronic, HIPAA's Security Rule requires reasonable administrative, technical and physical procedures to prevent unauthorized access, use, or disclosure of protected health information.[167] For example, in 2017, Metro Community Provider Network agreed to pay a $400,000 fine following a data breach in one of its hospitals that resulted from failures to secure patients' data and conduct a risk analysis, which would have revealed the glaring vulnerabilities in the hospital's data security program.[168]

For an AI system to be deployed in the healthcare context, the data that it is developed on must be HIPAA-compliant.[169] For example, when IBM engages with Memorial Sloan Kettering to provide Watson to help diagnose tumors, the data used to train Watson on how to identify tumors must have been gathered in compliance with HIPAA. However, HIPAA's approach to privacy fails to address the issue of de-anonymization, as made possible by computing power and massive quantities of data.[170] Consider an example from the mid-1990s to demonstrate the fact that even anonymous information stripped of identifiers can easily be de-anonymized to reveal individuals within datasets.[171] In Massachusetts, Group Insurance Company ("GIC"), a government agency, decided to release hospital records for every state employee in the government's health

165. 45 C.F.R. § 164.502(b) (2019).

166. REDHEAD, *supra* note 158, at 6.

167. Health Insurance Reform: Security Standards, 68 Fed. Reg. 8334 (Feb. 20, 2003).

168. Press Release, U.S. Dep't of Health & Human Servs., Overlooking Risks Leads to Breach, $400,000 Settlement (Apr. 12, 2017), https://www.hhs.gov/about/news/2017/04/12/overlooking-risks-leads-to-breach-settlement.html [https://perma.cc/8Q7U-DDEW].

169. *See generally* Abner Weintraub, *Consider HIPAA When Using AI & Machine Learning*, MEDSTACK (Nov. 14, 2017), https://medstack.co/blog/consider-hipaa-using-machine-learning/ [https://perma.cc/Q3W2-ZBZZ].

170. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1703–06 (2010).

171. *Id.* at 1719–20.

insurance program.[172] While names, addresses, and social security numbers were stripped from the data, patients' zip codes, birth dates, and sexes were not. [173] A graduate student purchased the voter rolls of Cambridge, Massachusetts (where the Governor of Massachusetts lived), which contained the names, addresses, zip code, birth date, and sex of every voter.[174] By combining the voter dataset with the hospital dataset, the student was able to identify the Governor's health records, including diagnoses and prescriptions.[175] This illustrates the problem of relying on anonymity as a total solution to ensuring health privacy — whenever an AI system is trained using multiple datasets, there is a very real possibility that the system will combine the datasets and arrive at a new base of knowledge that violates the requirements of HIPAA. One study concluded that 87% of the population can be uniquely identified merely with one's zip code, birthdate, and sex.[176] While other studies have found a slightly lower likelihood of identification,[177] that there is still any heightened chance of identification is problematic. It is vital to realize that the more data an AI system accumulates, the more likely it is that anonymous data can be de-anonymized.

Considering that datasets may be combined, transferred, or split up, AI systems that operate in the healthcare space should make the sources of their data transparent. This will give patients and other stakeholders the ability to determine whether anonymity is jeopardized and how medical data is used or transferred. Absent such transparency, we risk the possibility that others may be able to match multiple datasets, dissect the datapoints, and quickly pinpoint a particular individual who should have remained anonymous.

Two further issues bear special mention here. First, AI systems trained to diagnose illnesses and diseases require huge amounts of historical medical data, which is usually held by and is accessible through large firms and entities, whose interests may not align with the public's interests in welfare or privacy.[178] This misalignment may be amplified when the data was collected by nonprofit organizations, such as public hospitals, and transferred to for-profit entities.[179] In

172. *Id.* at 1719.
173. *Id.*
174. *Id.*
175. *Id.* at 1720.
176. *Id.* at 1719.
177. *Id.*
178. *See, e.g.*, Ornstein & Thomas, *supra* note 146.
179. *See, e.g.*, *id.*

these situations, the arrangement between the public seller and private buyer may contain exclusivity clauses, which would limit the ability of the public hospital to share the data with other actors (thereby potentially hurting public welfare), or grant the for-profit firm exclusive rights to determine when and how the data is transferred to third parties (thereby implicating privacy concerns).[180] These issues recently came to light following revelations of an agreement between the company Paige.AI and Memorial Sloan Kettering, regarding the use of millions of patient tissue slides, which Paige.AI would use to train AI systems to make diagnoses.[181] Through the use of this data, Paige.AI was able to obtain a market advantage over its competitors.[182]

b. AI for Children and Education

AI systems are not only used to diagnose diseases, hire employees, or predict creditworthiness. They are also increasingly used in toys and applications for children. Recently, a doll developed by Mattel that included an AI system was withdrawn from the market following privacy concerns regarding the effect it would have on child development.[183] Another doll outfitted with an AI system can apparently read children's emotions,[184] while other systems can tutor children.[185] Video games, a source of entertainment for millions of children, are rife with AI programs that can create characters, adjust the trajectory of the game to match the child's response, and shape the child's understanding of the world.[186] In millions of homes, AI

---

180. *See, e.g.*, *id.*

181. *Id.*

182. *Id.*

183. *See* Hayley Tsukayama, *Mattel Has Canceled Plans for a Kid-Focused AI Device that Drew Privacy Concerns*, WASH. POST (Oct. 4, 2017), https://www.washingtonpost.com/news/the-switch/wp/2017/10/04/mattel-has-an-ai-device-to-soothe-babies-experts-are-begging-them-not-to-sell-it/?utm_term=.8fee600e7aa1 [https://perma.cc/MB5U-ERYS].

184. *See* Timothy Revell, *Smart Doll Fitted with AI Chip Can Read Your Child's Emotions*, NEW SCIENTIST (June 20, 2017), https://www.newscientist.com/article/2137835-smart-doll-fitted-with-ai-chip-can-read-your-childs-emotions/ [https://perma.cc/5QZQ-74P3].

185. *See* Kumba Sennaar, *The Artificial Intelligence Tutor — The Current Possibilities of Smart Virtual Learning*, TECHEMERGENCE (Nov. 2, 2017), https://www.techemergence.com/artificial-intelligence-tutor-current-possibilities-smart-virtual-learning/ [https://perma.cc/TJU6-7KUN].

186. *See* Harbing Lou, *AI in Video Games: Toward a More Intelligent Game*, SCI. NEWS BLOG (Aug. 28, 2017), http://sitn.hms.harvard.edu/flash/2017/ai-video-games-toward-intelligent-game/ [https://perma.cc/XC3K-2QKV].

systems are on standby, waiting for a child to ask them a question.[187] Depending on the specific use of these AI applications, they may run afoul of privacy legislation that protects children. Further, as discussed above, the specific vulnerability in AI systems is that data accumulated over time may not have been collected in compliance with privacy statutes, creating a risk that data held and used by the AI programmers may come from an illegitimate source.

Congress has enacted numerous statutes that regulate data privacy when it concerns children and education, and the Transparency Model can help AI system stakeholders comply with them. The collection of data on children and the potential negative consequences of failing to protect children's identities are some of the main concerns that gave rise to the Fair Information Practice Principles.[188] The Children's Online Privacy Protection Act of 1998 (COPPA) prohibits the collection of data about children under thirteen without their parents' consent.[189] The onus is on AI data providers and trainers to ensure that the data they use to train AI systems was not collected from children without their parents' consent.[190] Moreover, data collectors must ensure that their data does not "run away" from them to third parties or unknown entities, as parents retain the right under COPPA to request destruction of the data collected on their child.[191] These requirements imply that AI trainers should be aware of how their datasets were collected, and on whom they have stored data.

COPPA created numerous obligations for multiple entities in the AI design process, including the AI system operators that create applications marketed towards children. Operators have to make the data they are collecting transparent, and must reveal how they collect

---

187. *See* Allison Aubrey & Michaeleen Doucleff, *Alexa, Are You Safe for My Kids?*, NAT'L PUB. RADIO (Oct. 30, 2017), https://www.npr.org/sections/health-shots/2017/10/30/559863326/alexa-are-you-safe-for-my-kids [https://perma.cc/5DUU-V6WV].

188. *See* F.T.C., PRIVACY ONLINE: A REPORT TO CONGRESS 4–6, 12–14 (June 1998), https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf [https://perma.cc/ZWL6-NQEY]. The Fair Information Practice Principles were early attempts to codify principles concerning the collection of information on children. *See generally id.*

189. *See* 15 U.S.C. §§ 6501–6506 (2012).

190. *See Children's Online Privacy Protection Rule: A Six-Step Compliance Plan for Your Business*, FED. TRADE COMM'N, https://www.ftc.gov/tips-advice/business-center/guidance/childrens-online-privacy-protection-rule-six-step-compliance [https://perma.cc/TWN5-7ZVH].

191. 16 C.F.R. § 312.6 (2019).

and use it.[192] Operators also have to disclose whether they have sold the dataset to other operators, and the parents would have to consent to that.[193] The proposed Transparency Model, therefore, fits squarely within the requirements of COPPA and would help operators in this space comply with their already existing legal obligations.

c. Data, AI, and Consumer Finance

Financial services is the leading economic sector adopting AI today, and it seems entities in this arena will continue to invest in AI moving forward.[194] Lenders and financial institutions utilize AI systems in their operations,[195] and the market is expanding so rapidly that the impact is already noticeable.[196] AI systems utilize social media platforms, handwriting style, and other unorthodox sources of data to predict whether an applicant would be a trustworthy borrower, among other things.[197] Financial institutions access loan histories, credit reports, and purchase histories that detail consumer behaviors on a granular level; AI systems incorporate these traditional and untraditional data sources in an attempt to predict the creditworthiness of potential borrowers.[198]

---

192. *Id.* § 312.4.

193. *See Children's Online Privacy Protection Rule: A Six-Step Compliance Plan for Your Business*, *supra* note 190.

194. *See* MCKINSEY GLOBAL INST., *supra* note 1, at 19.

195. Penny Crosman, *Can AI Be Programmed to Make Fair Lending Decisions?*, AM. BANKER (Sept. 27, 2016), https://www.americanbanker.com/news/can-ai-be-programmed-to-make-fair-lending-decisions [https://perma.cc/8ZKZ-YGCX].

196. *See* Calvert, *supra* note 3 (noting that AI can be used to predict opportunity in mergers and acquisitions, and discussing how AI "is influencing every aspect of the global economy, including investment banking"); DELOITTE, *supra* note 3, at 2 (concluding that most of the financial industry uses digital tools).

197. Crosman, *supra* note 195; *see also* Julius Adebayo & Mikella Hurley, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 202 (2016); Schellmann & Bellini, *supra* note 2.

198. BUSINESS CONSULTING INDUSTRY REPORT: JULY 2017, WELLS FARGO (July 2017), https://www.wellsfargo.com/com/securities/markets/equity-sales/prime-services/publications/update-07-17/#data [https://perma.cc/F9P4-TNAV] ("[Firms] collect and analyze big data and/or commission the data from the growing number of tech firms and data aggregators that process such data into useable reports for financial companies . . . . Sources include: meteorological and agricultural data; energy supplies and usage (e.g., oil tankers and storage levels); shipping/freight activity; construction activity; sensors from internet-connected machines or 'smart' devices (IoT sensors); pharmacological prescription data; e-commerce receipts and credit-card transaction data; government data; and retail brick and mortar activity (e.g., parking-lot photos). In addition, a large source of data consists of the information that web services and mobile apps already receive from users and the 'data exhaust' from many tech companies (e.g., Foursquare = GPS foot traffic), as

In short, AI systems not only track human behavior on- and off-line, but also reach all sorts of financial decisions — determining which loans consumers qualify for, what interest rates to charge, where to invest certain funds, and whether a particular transaction will be profitable.[199] One AI system claims to consider over twelve thousand different variables concerning a credit applicant.[200] One financial firm, Underwrite.AI, describes its use of machine learning:

> The focus of machine learning today is to create computer algorithms that learn from data and can make accurate predictions of [creditworthiness] based upon the patterns deduced within the data. Unlike traditional statistical modeling, the predictive models of machine learning are generated by the computer algorithm, as opposed to determinations made by statisticians based upon their interpretation of the results of linear regression and related techniques.[201]

The combination of exponential increases in data collection with the ability of AI systems to process massive amounts of data creates new legal challenges and risks for financial institutions.[202]

These emerging business practices must comply with the Fair Credit Reporting Act (FCRA) and other regulations protecting consumer borrowers. The FCRA requires consumer reporting

---

well as social media and social sentiment data, geolocation information, and online pricing and inventory data . . . While anonymized data can be provided to customers through contractual arrangements or an online services' application program interface (API), such data are also gathered through various methods, including aerial surveillance (e.g., microsatellites, drones and thermal imaging), beacons, and radio-frequency identification (RFID) sensors, and further analyzed using sophisticated software and AI deep-learning technology.").

199. Fortune Staff, *How AI Is Shaking up Banking and Wall Street*, FORTUNE (Oct. 22, 2018), http://fortune.com/2018/10/22/artificial-intelligence-ai-business-finance/ [https://perma.cc/29ZU-C7SE].

200. Jon Walker, *Artificial Intelligence Applications for Lending and Loan Management*, TECHEMERGENCE (Mar. 27, 2018), https://www.techemergence.com/artificial-intelligence-applications-lending-loan-management/ [https://perma.cc/CQG3-BRK2].

201. *About Us*, UNDERWRITE.AI, https://www.underwrite.ai/about [https://perma.cc/HW4P-W2UZ].

202. Nico Grant, *Hedge Funds, Beware: You Might Be Relying on Illegal Data Feeds*, BLOOMBERG (Dec. 7, 2017), https://www.bloomberg.com/news/articles/2017-12-07/hedge-funds-beware-you-might-be-relying-on-illegal-data-feeds [https://perma.cc/NQL3-GZX8] ("'[H]edge funds might get in trouble if they're buying data feeds from companies that didn't get consumers' approval to pass on information like credit card transactions, online browsing and emailed receipts,' said Jonathan Streeter, who worked at the U.S. Attorney's Office for the Southern District of New York under Preet Bharara. The U.S. Securities and Exchange Commission, in particular, could pursue enforcement cases.").

agencies to have procedures that ensure the credit process is "fair and equitable to the consumer, with regard to the confidentiality, accuracy, relevancy, and proper utilization of such information."[203] It also creates rights for consumers with respect to their credit reports, and places requirements on entities that collect, distribute and use these reports.[204] The FCRA's primary aim is to require that information about a credit applicant is accurate, and to ensure procedures that can ameliorate errors.[205]

Credit reports will typically include a massive amount of personally identifying information.[206] When a consumer alleges that an aspect of a credit report is inaccurate, the reporting agency must look into the consumer's claims.[207] Reporting agencies must also safeguard the reports by releasing them only for specifically enumerated purposes and by ensuring that the requester identifies him or herself, states the purpose for why the information is needed, and manifests that the information will only be used for permissible purposes.[208]

Many challenges arise concerning where to store and how to treat the underlying data that facilitates these financial decisions. Most AI systems that evaluate credit applicants offer third-party services, entailing some sort of exchange between the data owner and the data holder, and raising concerns about which entities have access to what data.[209] Of course, these third parties are responsible for maintaining the security of the data.[210] Still, the third party may utilize the data to improve its service and inform its operations for future applicants.

---

203. 15 U.S.C. § 1681(b) (2012).

204. 15 U.S.C. § 1681 (2012).

205. *Id.*

206. MARGARET MIKYUNG LEE, CONG. RESEARCH SERV., RL31666, FAIR CREDIT REPORTING ACT: RIGHTS AND RESPONSIBILITIES 6–7 (2013) ("[U]sually the individual's full name, Social Security number, address, telephone number, and spouse's name; financial status and employment information, including income, spouse's income, place, position, and tenure of employment, other sources of income, duration and income in former employment; credit history, including types of credit previously obtained, names of previous credit grants, extent of previous credit, and complete payment history; existing lines of credit, including payment habits and all outstanding obligations; public record information, including pertinent newspaper clippings, arrest and conviction records, bankruptcies, tax liens, and lawsuits; and finally a listing of bureau subscribers that have previously asked for a credit report on the individual.").

207. *Id.* at 7.

208. *Id.* at 7–8.

209. Kate Berry, *CFPB Catches Flak from Banks, Credit Unions on Risks of AI*, AM. BANKER (Dec. 6, 2018), https://www.americanbanker.com/news/cfpb-catches-flak-from-banks-credit-unions-on-risks-of-ai [https://perma.cc/R5TX-5MUQ].

210. *Id.*

For example, maybe it would create an internal profile for individuals, to identify that person across applications. This could lead to situations where an applicant is rejected not because of objective evidence, but because of previous determinations made by the AI system. In other words, it is conceivable that a negative determination for an individual based on a dataset could result in a negative determination for that individual permanently. If the AI system denies credit to someone who is actually creditworthy, and this mistake is not corrected, it will be reinforced over time, causing serious detriment to that person. The stakes involved thus support our proposal that companies involved in this space act very cautiously and ensure that people are not wrongfully excluded from the credit marketplace.

It is also of note that the implementing regulation of the Equal Credit Opportunity Act (ECOA) makes clear that "lenders who use non-traditional factors in making credit determinations must implement 'empirically derived, demonstrably and statistically sound, credit scoring systems.'"[211] Moreover, "[t]hese tools must be 'developed and validated using accepted statistical principles and methodology,' and must be subject to ongoing evaluation and review."[212] By its terms, the ECOA requires measures of transparency regarding the datasets used by AI systems. First, the statute requires a "demonstrably" acceptable model, so the lender must demonstrate that the underlying training conducted ensures the system's soundness.[213] More significantly, the statute requires the system to be "subject to ongoing evaluation and review."[214] Presumably, the statute would only require transparency with respect to regulators, and not necessarily the public at large. At a minimum, however, the statute prohibits complete black box machine learning systems that would make lending decisions without any accountability or oversight. Considering that an AI system could violate anti-discrimination law even while lacking any prejudicial mindset, it behooves trainers of AI systems to ensure that the data they use will not "taint" the AI system.[215]

---

211. David. C. Vladeck, *Consumer Protection in an Era of Big Data Analytics*, 42 OHIO N. U. L. REV. 493, 514 (2016).

212. *Id.*

213. *Id.*

214. *Id.*

215. Zarsky, *supra* note 99, at 1393–94.

2. General Normative Expectations of Privacy

In previous sections, this Article discussed how legislation addresses privacy and the kinds of requirements imposed on AI stakeholders who work in specific economic sectors. Privacy is a nebulous concept, and its precise definition and contours have been extensively debated.[216] This work does not purport to treat every detail of this complex topic. That said, this section of the Article aims to show that even when no statute prohibits a particular AI application because of privacy concerns, the industry can nonetheless violate expectations of privacy as they exist normatively, and can also erode those expectations over time as technology evolves. For example, in 2006, AOL released twenty million search queries of 650,000 of its users.[217] While it removed individuals' usernames and IP addresses, reporters were quickly able to identify User No. 4417749 based on the user's three search queries: "landscapers in Lilburn, Ga," a few people with the last name "Arnold," and "homes sold in shadow lake subdivision gwinnett county Georgia."[218] The reporters successfully identified a sixty-two-year-old widow named Thelma, who admitted she had also searched for topics such as "60 single men," and "dog that urinates on everything."[219] This demonstrates that even data that was "anonymized" may not actually have been so, or that this process may have occurred in an insufficient manner.

Anonymization becomes a problem when huge datasets are released to the public without regard to the implications of the Mosaic Theory.[220] The Mosaic Theory posits that even where the typical identifying characteristics of a person are obscured, such as name, social security number, date of birth, address, etc., in a large dataset one can de-anonymize someone probabilistically because of

---

216. Patricia Sanchez Abril, *Selling Privacy*, 2014 ANNUAL ACADEMY OF LEGAL STUDIES IN BUSINESS (ALSB) CONFERENCE, https://www.alsb.org/wp-content/uploads/2015/01/NP-2014-Selling-Privacy-Abril.pdf [https://perma.cc/HE6Q-GWLY] (understanding privacy through copyright regimes and the feasibility of selling personal data); *see also* Joseph W. Jerome, *Buying and Selling Privacy Big Data's Different Burdens and Benefits*, 66 STAN. L. REV. ONLINE 47, 47 (2013) (arguing that big data is transforming individual privacy — and not in equal ways for all, thus addressing the possibilities of buying and selling data).

217. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1717 (2010).

218. *Id.* at 1717–18.

219. *Id.*

220. *See generally* Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311 (2012).

the minimal likelihood that someone else shares the combination of a multitude of other characteristics.[221]

Widespread adoption of AI systems could increase the potential of de-anonymization. Presently, AI systems cannot function without vast amounts of data, as they must have a sufficient number of datapoints to effectively learn patterns and come to conclusions. In other words, the proliferation of AI systems automatically increases demand for large datasets, and the larger these datasets become, the easier and more likely de-anonymization becomes. With this growth in demand for big data, it is increasingly vital that AI operators fully disclose the sorts of data they use and the sources of that data. This would enable individuals to determine the various ways that data is being used in their favor or to their detriment. Considering that datasets are routinely transferred from one company to another, which may or may not be completely unrelated,[222] absent transparency it would be impossible for a consumer to know just how much their reputation precedes them when they travel from one website to another, one credit application to another, or one job to another.

3. Privacy by Design

Privacy by Design is the notion that throughout the design and lifecycle of a product or service, privacy should play a prominent role when considering various options.[223] There are seven "foundational principles" of Privacy by Design, the first of which stating that privacy initiatives should be "proactive not reactive, preventative not remedial."[224] In other words, privacy should be the default setting, embedded into the design itself throughout the product's entire lifecycle.[225] The product should have fully functional privacy

---

221. *Id.* at 320 ("The mosaic theory requires courts to apply the Fourth Amendment search doctrine to government conduct as a collective whole rather than in isolated steps . . . . [It] is therefore premised on aggregation: it considers whether a set of nonsearches aggregated together amount to a search because their collection and subsequent analysis creates a revealing mosaic.").

222. *See, e.g.*, AnnaMaria Andriotis & Emily Glazer, *Facebook & Financial Firms Tussled for Years over Access to User Data*, WALL ST. J. (Sept. 18, 2018), https://www.wsj.com/articles/facebook-sought-access-to-financial-firms-customer-data-1537263000 [https://perma.cc/BBE8-LLUT] (reporting that Facebook received financial data from some users who communicated with their banking institutions via Facebook Messenger).

223. Eric Everson, *Privacy by Design: Taking Ctrl of Big Data*, 65 CLEV. ST. L. REV. 27, 28 (2016).

224. *Id.*

225. *Id.* at 31.

protections with end to end security, and should be continuously visible and transparent regarding user privacy.[226]

Privacy by Design incorporates the "fair information practices common in most privacy legislation in use today: notice, choice and consent, proximity and locality, anonymity and pseudonymity, security, and access and recourse."[227] Privacy by Design offers an efficient framework for an AI developer to ensure that the data used to train an AI system is not compromised. Even though Privacy by Design focuses on visibility and transparency, "there should not be any confusion around the necessity to encrypt, obscure, or otherwise properly protect data."[228] Whether traditional personally identifying data, user data, or other potentially sensitive data, there is an inherent duty to safeguard the privacy interest in the data collected, stored, or used.[229]

For many years, the FTC has advocated for Privacy by Design:

> "Privacy by Design" and "Security by Design" [are] both concepts [that] seek to protect consumers' privacy and security from the outset of product design. In an era where AI systems and machine learning, algorithms and big data sets can hire and fire, inform health care decisions and extend financial opportunities, it is vital that these technologies do not run counter to established legal protections or public policy goals. In the same way companies incorporate privacy and security, so too should we have Data Ethics by Design.[230]

The concern for AI system operators is that the data they use to train the system not run afoul of privacy requirements and expectations. As discussed, statutes require special treatment for information pertaining to health, children, and consumer finance, and normative privacy expectations also place some limits on the extent to which an AI system will use datasets that violate consumer privacy.

In summary, using specific types of data, or using data in specific contexts, can create issues that implicate privacy concerns that trainers of AI systems ought to consider. Whether it is the type of data, such as health data, or the subject of data, such as children, or the context in which data is used, such as in consumer lending or

---

226. Ari Ezra Waldman, *Privacy's Law of Design*, UC IRVINE L. REV. (forthcoming 2019) (on file with authors).

227. Everson, *supra* note 223, at 29.

228. *Id.* at 33–34.

229. *Id.* at 34.

230. BIG DATA: INDIVIDUAL RIGHTS AND SMART ENFORCEMENT, F.T.C., 2016 WL 5791537, at *5.

other economic sectors protected by anti-discrimination legislation, AI systems are operating in spaces governed by existing privacy and anti-discrimination regulatory regimes. Hence, AI stakeholders, including operators, must ensure that the datasets used to train their AI systems do not violate these regulatory regimes. One way to start meeting this obligation is through the Transparency Model proposed in the following Part of the Article.

III. THE AI DATA TRANSPARENCY MODEL

A. The Need for an AI Data Transparency Model

The AI Data Transparency Model is a first step towards ensuring that the data used to train AI systems complies with all relevant regulations and societal expectations, which may otherwise limit the AI's use. In previous sections, this Article has identified some of the many risks that AI systems could pose for individuals and society as a whole.[231] This work also discussed the myriad ways in which the data used in AI systems could run afoul of privacy and anti-discrimination legal regimes.[232] The Transparency Model argues that dataset users up and down the data supply chain must ensure that the data remains in compliance with existing laws. In cases where AI systems are provided or exposed to data, the trainers should actively ensure that the providers include assurances of the data's propriety. Audits should evaluate the sources of the data, the data's contents, and whether the data's use complies with any regulatory limits that arise. These procedures should be reasonable and flexible, so that they can conform to the type of data used and the role that the AI system plays. Considering the crucial role that data plays in creating AI systems and their outcomes, these procedures should be adopted and standardized by stakeholders. There should also be a process whereby conformity with the standards proposed herein can be evaluated and certified by objective third parties. This will incentivize best practices by encouraging transparent operations, which would increase reputational risks for developers that do not pay appropriate attention to the dangers enumerated above. While the suggested Transparency Model will not solve all of the emerging issues associated with AI systems, it will help clarify some of them while also illuminating further areas of concern and giving guidelines to an industry that currently operates in a regulatory vacuum.

---

231. *See supra* Part II.
232. *See supra* Section II.B.2.

There are four main components to the proposed AI Data Transparency Model. First, stakeholders in the AI systems industry should conduct audits and examine the data their AI systems are exposed to, according to the type of data collected and the risk of misuse. Second, stakeholders should be required to retain the data they used to train the AI in case there is a future need to further scrutinize how the AI system was developed. Third, audits should be standardized and conducted by objective third parties who are not the developers themselves; these third parties should also certify the results of the audits. Finally, the Transparency Model provides for a safe harbor — AI systems operators should enjoy some protections from liability in limited circumstances, when they earnestly comply with the Model but harm occurs nonetheless.

The four components of the Transparency Model are just some of many possible responses to the AI revolution. One alternative would be to simply do nothing and allow this advanced technology to develop further before trying to rein it in. But this alternative is a dangerous one, as it does nothing to incentivize the cultivation of best practices that consider the public interest early on.[233] Another alternative is to reject the idea that individuals should have privacy interests in their own data. After all, each individual's respective, tiny piece of data constitutes a minuscule percentage of the massive datasets out there, composed of billions of individuals (as the adage goes, the whole is greater than the sum of its parts). At the other extreme, perhaps data should be controlled much more stringently, by government regulators who would mandate inspections and require disclosures, akin to the Securities and Exchange Commission or the Internal Revenue Service. The proposed Transparency Model represents an appropriate balance between these two extremes, and will help incentivize more ethical uses of data by AI developers without threatening or hindering the development of the technology altogether.

The most prudent way to balance the need for AI innovation against the dangers of discrimination, privacy violations and other transgressions such as copyright infringement, is to mandate and encourage disclosures of the types of data used and the manner in which this data helps AI systems produce their output. The disclosures should be the result of scrutinizing audits of the data, which ought to evaluate the data's integrity, origin, and quality, as well as identify any potential for violating discrimination or privacy

233. Waldman, *supra* note 226.

statutes and norms.[234] These periodic audits should evaluate both the training phase of the AI system, and how the system is continuously operated.[235] They should examine where the datasets came from, what information they contain, and what permissions and limitations affect how the datasets can be used.[236]

With all this information collected, the auditor should look at what the AI system does and verify that its particular use, combined with the type of data, does not run afoul of any regulatory requirements.[237] It is imperative that the auditor be someone other than the stakeholders themselves. Depending upon the sort of data collected and the AI system's purpose, this audit should examine the data according to existing legal rules, such as those described above. Problems resulting from bad data and harmful AI systems are not necessarily attributable to a lack of legislation that proscribes negative outcomes; rather, they are the result of the sophisticated manner in which AI systems operate, which eludes easy determinations of whether such practices comply with existing legislation. Further, the difficulty of enforcing these laws on AI systems is partially due to a lack of understanding of how AI systems work, as well as an overreliance on, and false overconfidence in, advanced technologies.[238]

The purpose of the auditing process is to ensure reliability and trust in the AI industry, while also enabling the technology to keep growing and developing. The Transparency Model seeks to establish a set of standards that would help address and reduce some of the most common pitfalls and issues concerning misuse of datasets in AI systems. The Model can be compared to the NIST Framework as promulgated by the National Institute of Standards and Technology, which created a multi-step process an entity can take to determine its cybersecurity and data protection risks as well as to evaluate

---

234. Chander, *supra* note 102, at 1025 ("[T]he problem is not the black box . . . but the real world on which it operates. We must design our algorithms for a world permeated with the legacy of discriminations past and the reality of demonstrations present.").

235. Lehr & Ohm, *supra* note 45, at 655 (discussing the eight steps of an AI system's development).

236. *See* Mattioli, *supra* note 38, at 536–37 (discussing how data changes hands).

237. *See supra* Part II.

238. *See generally* BRUNDAGE ET AL., *supra* note 23, at 51 (discussing the rapidly growing capabilities of AI and machine learning while highlighting the under-focused collateral consequences from cyberthreats).

appropriate responses.[239] The AI Data Transparency Model consists of a set of checklists and tasks that provide a roadmap for AI developers, trainers and operators to ensure that the data being used complies with all relevant requirements. In this sense, the Transparency Model works to ensure that privacy and equality protections are considered at each stage of AI development, rather than only as an afterthought.

Another part of the Transparency Model will be to certify that AI systems were successfully audited for unreasonable risks to privacy and discrimination. This part of the Model is inspired by other instances of disclosure and transparency rules that have worked in other industries, such as food and drugs, and is consistent with other scholars' approaches to these issues.[240] There are many examples of certification programs that indicate whether a product has successfully met a particular set of standards: the organic food labeling regime, which has helped foster a market for organic foods and educates consumers about the ingredients in their foods;[241] or the Kosher label added to foods after inspection by the relevant authorities.[242]

Finally, where an AI stakeholder substantially complies with the Transparency Model, they should enjoy some degree of safe harbor against liability for limited and inadvertent mistakes made by an AI system. The Model does not include a strict liability regime for all violations resulting from an errant AI system or slight mistakes in datasets. This safe harbor serves numerous purposes. First, it preserves judicial economy by preventing plaintiffs from insisting retroactively on discovery into massive datasets. This is particularly important when violations might be nothing more than inadvertent infringements that represent small deviations from regulatory requirements, in contrast to harms that result from an unwillingness to protect against foreseeable consequences that derive from lack of attention to the dataset's composition and origin. Second, the Model

---

239. *See Cybersecurity Framework*, NAT'L INST. STANDARDS & TECH., https://www.nist.gov/cyberframework [https://perma.cc/4TU7-6BUZ].

240. For an argument that algorithms should be regulated in a similar manner to environmental impact statements, see generally Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2018).

241. *See generally History of Organic Farming in the United States*, SUSTAINABLE AGRIC. RES. & EDUC., https://www.sare.org/Learning-Center/Bulletins/Transitioning-to-Organic-Production/Text-Version/History-of-Organic-Farming-in-the-United-States [https://perma.cc/JDY9-NSZ5].

242. *See What Does Kosher Mean?*, BADATZ IGUD RABBONIM, http://www.koshercertification.org.uk/whatdoe.html [https://perma.cc/748Q-HQLS].

rewards desirable behavior by encouraging dataset owners to inquire into the source of what they are using to train their AI systems, and fosters sincere efforts to avoid violating the law.[243] Self-regulation works when it can incentivize the internalization of risk, rather than leaving consumers to accept the costs. The safe harbor would not protect against intentional infringement or negligent standards regarding the evaluation of datasets. As long as it is cheaper for an AI operator to implement safeguards than it would be to risk legal exposure, the operator would act rationally and seek to comply with the requirements of the safe harbor. Finally, establishing this juxtaposition will leave consumers better off, because the potential harms of bad data and misused datasets will be alleviated such that consumers will not be harmed in the first place.

B. The Benefits of the AI Data Transparency Model

1. The Benefit of Increased Transparency

The sort of transparency proposed in the AI Data Transparency Model is not only beneficial by preventing the harms discussed in Part I, regarding discrimination and privacy violations, but it is also beneficial as an end itself. By cultivating a culture of transparency, the fears and stigmas associated with AI systems can be mitigated, and the Transparency Model will help instill a trust that machine learning technologies will not be misused.[244] Transparency enables "shaming" in the event that an AI system produces inappropriate outcomes, though that presumes that there is an audience before whom to shame the wrongdoer, and that the operator in charge of the system would respond to such shaming.[245] Where an AI system functions inappropriately — by, for example, misidentifying people of color as criminals or rejecting great applicants for a job or educational opportunity because of their gender — transparency helps society understand what went wrong. Moreover, where the AI system operator is dedicated to implementing steps to prevent such

---

243. For a proposal that copyright liability for webhosting providers should be grounded in a safe harbor proposal, see generally Lital Helman & Gideon Parchomovsky, *The Best Available Technology Standard*, 111 COLUM. L. REV. 1194, 1217 (2011).

244. *See, e.g.*, Maureen Dowd, *Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse*, VANITY FAIR (Apr. 2017), https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x [https://perma.cc/FR2L-WCJ6].

245. Tal Z. Zarsky, *Transparent Predictions*, U. ILL. L. REV. 1503, 1534–35 (2013).

unfortunate outcomes, transparency helps that system operator show that this mistake was unfortunate, rectifiable, and not the result of wantonness or recklessness.[246] Further, transparency can help identify which actor in the data supply chain is responsible — is the violation the result of the data used to train the AI system, the biases of the AI's trainer, or an example of an AI system going out of tune and running amok? Additionally, many of the prescriptions offered here amount to "good press" as long as the AI operator is sufficiently in compliance, as evidence of such compliance can help foster trust between the operator and consumers.[247] By cultivating these positive industry practices and standards, transparency increases trust in machine learning technology itself and helps consumers understand the ways in which they benefit from the technology.

It is true that what "transparency" calls for exactly is often unclear,[248] but this is a strength of the Model proposed here, not a flaw. "Transparency" could mean having a particular person understand what aspects of their digital background led to a harmful AI result. It might mean that individuals ought to know the backgrounds of datapoints that are most similar to them, or the ways in which they are similar to datapoints within the underlying dataset. It could mean knowing the confidence level that the AI operator has in the system, or the error rate of that system. It can mean an understanding of the system itself, involving disclosure of information regarding the system's setup, the data used to train it, its objective and predictive success, and other aspects of its operation.[249] Under the Transparency Model, the meaning of "transparency" depends on the nature of the AI program being evaluated. When "transparency" serves as a means to achieve specific regulatory or business-oriented objectives, such as anti-discrimination, privacy protection, or some other concern that the dataset implicates, the Model can remain relevant and impactful across industries and AI systems. By flexibly adjusting to specific contexts, the Model helps avoid the problem of untailored or overly-tailored approaches, which are either too comprehensive for certain needs or too basic for others.

---

246. *See, e.g.*, Lee, *supra* note 121.

247. *See* Desai & Kroll, *supra* note 28, at 58, 67.

248. *See* Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 58 (2017).

249. *Id.* at 55–56.

2. Value Adding

At the core of the Transparency Model is the contention that "bad" data, whether discriminatory or privacy-violating, reduces the likelihood that AI systems will produce good outcomes.[250] By facilitating a framework that reduces the use of problematic data, the likelihood of adverse outcomes will similarly decrease. It is true that requiring compliance with the Transparency Model will likely increase the initial costs of training AI systems, which would inevitably raise costs for consumers of goods or services provided by or utilizing AI. However, the decreased risk of adverse outcomes could offset some of those increased costs, and once consumers gain trust in AI systems those costs would be offset even further. Moreover, creditors and investors can gain confidence in the AI system's development, which might result in lower interest expenses and transaction costs associated with raising capital. By certifying the provenance and legitimacy of data, AI developers can aid their investors in conducting due diligence.

Examples given throughout this Article provide support for this argument.[251] Unjustified denials of promotions in the employment context leads to talented individuals going unrewarded. This creates costs for the company in that it now has a worse employment hierarchy than it would have had, and in that it could lose that employee entirely because she feels slighted by the lack of promotion. Similarly, unwarranted denials of credit by AI systems that wrongly predict creditworthiness lead to less customers for credit card and finance companies, which leads to less revenue for that company and diminished economic activity for society at large. In the same vein, schools that use AI systems to evaluate student applications could see a weaker student body if that system fails to identify worthy applicants simply because they do not fit the "traditional" standard that system was taught. In each of these examples, the goal is clear: to find the best applicant. A more fine-tuned AI system can reach better results than a system whose operations and development are not scrutinized. Hence, while implementing the Transparency Model may increase costs due to auditing the data for all stakeholders in an

250. *See generally* Thomas C. Redman, *Bad Data Costs the U.S. $3 Trillion Per Year*, HARV. BUS. REV. (Sept. 22, 2016), https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year [https://perma.cc/Z4RB-SRQH] (noting that bad data costs U.S. corporations over $3 trillion a year, according to IBM).

251. *See supra* Part II.

AI system, from development to operations, the final output of those systems may very well be far more valuable.

In Part II above, this Article addressed possible evils that can result from faultily-composed datasets, which can lead AI systems to discriminate against protected groups or to violate privacy protections. Just because these injuries occur "beneath the surface," embedded within the training of an AI system, does not mean that they are excusable. As AI systems participate more and more in business operations, their developers will have to grapple with the serious issues that surround the ways AI systems use data. There are already examples of businesses that suffer reputational harms when datasets cause AI systems to perform inappropriately or illegally, and there is no sound reason to de-emphasize scrutiny in this space merely because the mechanisms are more complicated than traditional business practices.

3. Flexibility

As briefly touched upon above, the Transparency Model is flexible in nature. The scrutiny of the auditing process should change depending upon what is reasonable under the circumstances — in light of how the data was collected and what the responsibility of the trained AI will be. In other words, scrutiny may be higher or lower depending on the particular outcome that the AI system produces:

> If the private sector industry in question is regulated, such as the auto or pharmaceutical industry, evidence of correctness can naturally become a requirement. If the sector in question is not regulated . . . [the entity] should consider their requirements for demonstrating the correctness of their automated decision making as a best practice, since it will enable trust in their products and services.[252]

For example, an AI that drives cars, analyzes medical symptoms or manages pension fund portfolios should have its data more closely scrutinized than an AI created to play a board game.[253] These considerations inform the evaluator of what the probability of an undesirable outcome may be, and what the consequences for such an

---

252. *See* Desai & Kroll, *supra* note 28, at 45–46.

253. *Compare* Bernard Marr, *The Amazing Ways Tesla Is Using Artificial Intelligence and Big Data*, FORBES (Jan. 8, 2018), https://www.forbes.com/sites/bernardmarr/2018/01/08/the-amazing-ways-tesla-is-using-artificial-intelligence-and-big-data/#fb9fc8d42704 [https://perma.cc/UV5V-6A47], *with Google AI Defeats Human Go Champion*, BBC (May 25, 2017), https://www.bbc.com/news/technology-40042581 [https://perma.cc/6Z9E-8ATT].

outcome would look like. Regulated sectors, such as housing, employment, and healthcare must obviously comply with all relevant legal requirements, which would inform the evaluator about the harms she is scrutinizing the AI system for. On the other hand, where the AI exists for entertainment, the evaluator may consider the harms that privacy encroachments pose for the company's reputation, the need to collect such data, or the risk of harms that a breach of the data would cause the subjects in the dataset. In short, the evaluator would need to consider both legal requirements, where applicable, as well as normative and reputational ones, and have the flexibility to weigh AI data violations differently according to the circumstances.

Because AI is rapidly developing, maintaining a flexible approach to regulation can efficiently and fairly balance the utility of developing the technology against the costs that society could suffer.[254] Within the copyright regime, for example, a "technological fair use" doctrine would properly balance the equities between the need to grow the technology sector and the rights of the copyright holders to avoid AI systems from using and relying on their works.[255] This notion of fair use could also apply to the harms addressed here, to instances where the AI operators make clear and concerted efforts to avoid violative results that infringe on privacy or produce undesirable outcomes. In short, it is not the case that the nature of AI as a black box will inevitably lead to rampant discrimination, but its use will not lead to the end of discrimination either. It is the case, however, that encouraging AI operators to be transparent about how they use data and what that data is, and certifying that they are not behaving in ways that increase discriminatory outcomes, could steer AI systems towards less discriminatory outcomes, benefiting society at large.

The proposed Transparency Model focuses on what is ascertainable by inspecting and evaluating the data, rather than the mysterious inner-workings of an algorithm that returns results with unintelligible rationales.[256] It also helps fill a void, also noticed by other scholars, by focusing on data rather than on the algorithm that

254. Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 802 (2010) (arguing for a reconfiguration of the fair use defense, in copyright infringement cases, to encapsulate new uses of copyrighted works by technology).

255. *Id.* at 838.

256. *See* W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 432–34 (2015).

is using this data.[257] A rigid regulatory scheme risks both over-regulation and under-regulation. The former of these would subject certain AI systems to requirements that are unnecessary to achieve the regulation's objective of careful scrutiny of potentially harmful data. The latter would fail to protect against a developer or AI system that fails to proscribe undesirable data uses. Thus, the Transparency Model is intentionally limited in the sense that it does not have specific requirements that should apply to every AI system, apart from continued compliance with existing laws. AI systems exist in far-flung industries, utilizing vastly different datasets for any number of distinct uses — the solution devised to promote accuracy in results from AI systems must be equally flexible.

C. Theoretical Justifications

1. Law and Economic Theory: Transparency, Accountability, and Efficiency

At the core of the Transparency Model is the contention that transparency is a precursor to accountability: "If law and due process are absent from this field, we are essentially paving the way to a new feudal order of unaccountable reputational intermediates."[258] This means that society cannot hold a malicious algorithm accountable if it does not subject it to some measure of transparency. The converse of this is also true, that society must keep transparent the process of holding algorithms accountable. As one set of authors noted:

> Strong arguments support the position that algorithmic agents that operate without proper, or flawed, human oversight; or absent of well-defined governance and ethical frameworks, may have negative effects on greater societal norms and values such as the holy triumvirate of *liberte, egalite, fraternité* — or to put it in the language of the existing legal frameworks, fundamental human rights and freedoms, equality and social cohesion.[259]

Because these values are so integral to society, the implication is that transparency serves dual purposes — as a corollary norm to

---

257. Lehr & Ohm, *supra* note 45, at 655 ("[A]lmost all of the significant legal scholarship to date has focused on the implications of the running model . . . and has neglected most of the possibilities and pitfalls of playing with the data.").

258. Citron & Pasquale, *supra* note 3, at 19.

259. B. Bodo et al., *Tackling the Algorithmic Control Crisis — The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents*, 19 YALE J.L. & TECH. 133, 137 (2017).

fundamental human rights, and as a means to ensure that those fundamental rights are not infringed upon.

The theory of law and economics suggests that where the operations and development of AI systems are made more transparent, transactions between their operators and consumers will become more efficient.[260] Without transparency, there is no way for consumers to determine how much data collectors take, what sort of data they take, with whom they share it, and how they use it.[261] Transparency is a means by which society can verify that the drawbacks of AI systems do not outweigh the benefits. Because rights and responsibilities can conflict with one another, in order to ascertain whether an AI system imposes itself on other arenas that merit protection, some aspects of the system should be reviewable and accessible. Mandating transparency can also facilitate innovation by opening access to the data inputs used by one successful enterprise to the benefit of all. Hence, when an entity uses data in predictive analytics, but did not collect that data itself, a degree of transparency is necessary to ensure that contractual or statutory limitations on the use of that data are not breached.[262] Transparency can help alleviate the numerous risks created when this kind of data is used, including the possibility of security breaches, inadvertent nonconsensual uses, or worse — that a machine learning algorithm could use data in ways that violate privacy and anti-discrimination laws.

2. The Market Structure and the Multi-Player Model

A common characteristic of a completed AI system is that numerous entities contributed to its development. At least ten different types of actors can help program an AI system, including the software trainers, the data provider, the feedback provider (or trainer), the user, the owner and her employees, her investors, the public or the government, and even the AI itself.[263] Any of these players can expose an AI system to data, which can then be distributed to other players who may combine it with more data, resulting in new sources and insights.[264] This distribution of labor creates problems of knowledge and oversight over the contents of the

---

260. *See* Posner, *supra* note 130, at 394–403.

261. *See* Everson, *supra* note 223, at 33–34.

262. Max N. Helveston, *Consumer Protection in the Age of Big Data*, 93 WASH. U. L. REV. 859, 873–74 (2016).

263. *See* Yanisky-Ravid, *supra* note 29, at 692.

264. For a discussion of the Multiplayer Model, see generally Yanisky-Ravid & Liu, *supra* note 12, at 2216.

dataset. Because the sources of access to new data are so diffuse, it becomes more difficult to determine who should be responsible for bad data. Someone who downloads a dataset of millions of images to train a facial recognition AI may not realize that the dataset was biased in ways that could produce discriminatory outcomes. She may not realize, for example, that the datapoints therein are largely composed of Caucasian faces, limiting the utility of the dataset in recognizing other racial groups. This hypothetical demonstrates that the Multi-Player nature of AI development increases the risk that datasets could be compromised by impermissible datapoints.

Data is not only collected by entities that use it for their own purposes, but it is also distributed to other entities who may use it for the same, or for a different purpose than the original data collector.[265] Indeed, some data collectors retain data solely to distribute it to other users. Whenever datasets change hands, are adjusted for analysis, or combined with other data, the possibility of misuse or corruption grows.[266] This also complicates any effort to ensure consensual use of data, because at a certain point it could very well become impossible to ascertain where the specific datapoint originated from, and whether that datapoint was permissibly collected. A given consumer, concerned about the diffusion of their information, cannot readily control their information absent some degree of transparency. Additionally, even if the consumer initially consented to various piecemeal collections of their personal data, if those datapoints are aggregated the result could create a profile of that consumer far beyond anything she ever thought was possible. Thus, the main purpose of the AI Data Transparency Model is to focus on the long run, by incentivizing stakeholders to use reliable data, thereby focusing on the prevention of harm rather than the assignment of liability.

3. Law and Economic Theory: Self-Regulating Incentive Mechanism

One glaring problem of AI applications is how difficult it is for someone to know if they were incorrectly "rejected" or labeled "unworthy." For a host of reasons, self-regulation offers a more efficient route to ascertaining the chances that these unfortunate results will occur. First, the applicant or consumer may not be aware that the company is utilizing an AI system, so it would be impossible

265. *See* Helveston, *supra* note 262, at 874.
266. *See id.*

for her to suspect that her rejection occurred because of a misconfigured AI system. Next, considering the explosion of uses of AI systems in a wide variety of commercial arenas, it is unlikely that a government response could adequately regulate a technological tool that is used for all sorts of purposes in a multitude of economic sectors, transcending the regulatory jurisdictions currently carved out in the United States. Instead, each company that utilizes an AI system is in a far better position to make sure that it is operating within the confines of the law. Just as regulatory authorities depend on independent audits by accountants to ensure the financial propriety of large corporations, so too could the government depend on private sector audits that ensure compliance with federal law in AI systems. This is not to say that the government should play no role in regulating AI, but rather to argue that any effective regulatory regime will require private sector involvement and cooperation, and that all of this begins with transparency.

CONCLUSION

The advent of huge data sources facilitates numerous applications: "trends/pattern analysis; regulatory compliance; fraud detection and prevention; predictive analysis and modeling; incident prediction; geo-correlation; sentiment analysis; diagnostic and medical use; and others."[267] In some cases, the data used in machine learning tends to have numerous features in common. First, it is non-exclusive and non-rivalrous, and multiple parties can use the same data without consequence. For the most part, it is inexpensive and easily collected. Moreover, it is everywhere.[268] This Article demonstrates that AI systems are delving more and more into areas that are protected, regulated, and valued in special ways, implicating privacy and equality concerns. As AI systems delve into these highly sensitive areas of society, operators must carefully consider the risks that arise when they mistreat and misuse data. The AI Data Transparency Model suggested here can help address these risks. As part of the Transparency Model, AI stakeholders should implement privacy and equality by design, adopting guidelines that meet legal requirements from the very first moment they expose their system to data. Stakeholders should audit the datasets they use to train their AI, relying on independent third parties who can verify that their data

267. *Id.* at 870.

268. *See* D. Daniel Sokol & Roisin Comerford, *Antitrust and Regulating Big Data*, 23 GEO. MASON L. REV. 1129, 1136–37 (2016).

practices do not run afoul of existing norms and legislation. Additionally, a third-party certification mechanism can increase public trust regarding the use of AI systems, and help companies save face when mistakes do occur. Finally, to incentivize compliance with these recommendations, a safe harbor is appropriate for operators who make special efforts to avoid the dataset problems discussed. These recommendations will not be possible without significant increases in transparency, which is a means to facilitate continued innovation in this field without increasing the risk of public backlash. AI technologies have great potential to tremendously benefit society, but without transparent, careful progress, they can violate fundamental principles of equality, fairness, and privacy.