THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME XXV STLR.ORG SPRING 2024

BEYOND ALGORITHMIC DISCLOSURE FOR AI

Christopher S. Yoo*

One of the most commonly recommended policy interventions with respect to algorithms in general and artificial intelligence ("AI") systems in particular is the need for greater transparency, often focusing on the disclosure of the variables employed by the algorithm and the weights given to those variables. This Essay argues that any meaningful transparency regime must provide information on other critical dimensions as well. For example, any transparency regime must also include key information about the data on which the algorithm was trained, including its source, scope, quality, and inner correlations, subject to constraints imposed by copyright, privacy, and cybersecurity law. Disclosures about pre-release testing also play a critical role in understanding an AI system's robustness and its susceptibility to specification gaming. Finally, the fact that AI, like all complex systems, tends to exhibit emergent phenomena, such as proxy discrimination, interactions among multiple agents, the impact of adverse environments, and the well-known tendency of generative AI to hallucinate, makes ongoing post-release evaluation a critical component of any system of AI transparency.

* Imasogie Professor of Law and Technology, Professor of Communication, Professor of Computer & Information Science, and Founding Director of the Center for Technology, Innovation & Competition, University of Pennsylvania. Thanks to Timothy van Dulm for his top-notch library support and Simon Roling for his expert research assistance.

I. INTRODUCTION

One of the most striking developments of the past decade is the increasingly widespread use of algorithmic decision making, particularly in the context of artificial intelligence ("AI"). Advancements in algorithms and AI have improved business processes and radically expanded the capabilities available to consumers. The meteoric rise of generative AI and, in particular, the furor surrounding the release of ChatGPT-4 have raised the level of interest to new heights.

Increasing reliance on AI has inevitably expanded the need for understanding its limitations and potential deficiencies. These include its tendency to replicate biases that exist in the data on which it is trained, its ability to produce and amplify harmful content, the potential to impair individuals' privacy and security, the danger that users may become overly reliant on AI's outputs, and, in the case of generative AI, the possibility that it can create fictitious content (so-called AI hallucinations), among others.[1]

A common starting point in the legal academy for discussions about algorithmic accountability is Frank Pasquale's warnings about the dangers of "black box" algorithms, whose internal workings remain obscure.[2] For Pasquale, the key first step is the adoption of measures to make algorithms more transparent and intelligible.[3] Consistent with Pasquale's recommendation, a wide range of governmental authorities have issued or joined policy statements calling for AI to become more transparent and explainable.[4] Unfortunately, the authors of these

[1] Fui-Hoon Nah et al., *Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration*, 25 J. INFO. TECH. CASE & APPLICATION RSCH. 277, 284-88 (2023).

[2] For Pasquale's first invocation of the black box metaphor, *see* Frank Pasquale, *Battling Black Boxes*, MADISONIAN (Sept. 21, 2006), https://madisonian.net/2006/09/21/battling-black-boxes/ [https://perma.cc/2PQW-M67P]. The metaphor eventually became the centerpiece of his book on algorithmic regulation. FRANK PASQUALE, THE BLACK BOX SOCIETY 2-3 (2015).

[3] PASQUALE, *supra* note 2, at 141-42; Frank Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, 104 NW. U. L. REV. 105, 109 (2010). For another early article laying out a conceptual justification for transparency that also acknowledges its limitations, *see* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1504.

[4] *See, e.g.*, WHITE HOUSE OFF. OF SCI. & TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE 44, 51 (2022),

statements have pitched them at such a high level of generality that they provide little practical guidance as to their implementation.

This Essay explores the limitations of disclosures about algorithms and examines additional dimensions along which algorithmic transparency might be implemented. Part II discusses the limitations of disclosing algorithms and argues that bare code may not be revealing enough and that inclusion of variables generally thought to be problematic may reflect a desire to remedy flaws in the training data. Part III discusses the importance of disclosures about the data on which AI models were trained and examines how different aspects about data quality should be disclosed. Part IV discusses the importance of disclosing information about the ways in which algorithms are validated and tested. Part V explores how complex systems typically give rise to emergent behavior that appears after deployment at scale and requires post hoc assessments of how the algorithm is behaving in the real world. Part VI concludes.

II. DISCLOSURES ABOUT ALGORITHMS

Perhaps the most basic form of transparency would simply require developers to disclose the code comprising their algorithms either to regulators or to the public.[5] This approach implicitly presumes that examining the variables that an algorithm takes into account and the weights that it gives them is sufficient to allow people to understand its likely behavior. At a minimum, algorithmic disclosure

---

https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [https://perma.cc/RDH5-FLZH]; NAT'L INST. OF STDS. & TECH., U.S. DEP'T OF COM., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) 3, 12-13, 15-16 (NIST AI 100-1, Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [https://perma.cc/DE7P-FNV3]; *European Declaration on Digital Rights and Principles for the Digital Decade* (EU), 2023 O.J. (C 23) 1, 5, paras. 9(b), 10, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023C0123(01) [https://perma.cc/GWC3-XBWC]; Org. for Econ. Coop. & Dev., *Recommendation of the Council on Artificial Intelligence* § 1.3 (OECD/LEGAL/0449, May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [https://perma.cc/J7UP-PKH7]; *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023* (Policy Paper Nov. 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [https://perma.cc/W6CU-TBYA]; *see also Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF [https://perma.cc/929N-JVVH] [hereinafter EU AI Act Proposal] (proposing an approach that varies the level of transparency required with the level of risk posed by the AI system).

[5] *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1308-10 (2007) (public); Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 24-25 (2014) (regulators); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1039 (2017) (reviewing PASQUALE, *supra* note 2) ("Pasquale's metaphor of the 'black box' suggests that the solution to algorithmic ills is to open algorithms up for examination.").

would reveal whether the algorithm relies on impermissible criteria, such as race, gender, or religion.

Commentators have questioned whether algorithmic disclosure provides meaningful transparency. As an initial matter, algorithms may be too complex for observers to understand fully.[6] This problem is even more protracted for concerns about algorithmic discrimination, in which the bias may be the result of bias in the training data that may not be discernable from decisions about which variables to include and the weights given to those variables.

Furthermore, transparency by itself will not reveal whether an algorithm has been able to rely on facially neutral proxies for variables deemed problematic.[7] The literature has identified examples of such reliance in such varied contexts as health care,[8] education,[9] and employment,[10] among others.

In addition, an algorithm attempting to correct for bias in training data may have to apply a correcting factor explicitly based on the prohibited variable.[11] Considering (and bolstering) vulnerable classes in a model can reduce discrepancies in accuracy resulting from biases included in the data used to train the model.

In short, simply knowing the code that comprises an algorithm may not provide a complete understanding of its real-world behavior. In addition, the inclusion of factors that may seem problematic on their face may be justifiable by the need to correct for other factors.

Furthermore, algorithmic disclosure must confront certain legal constraints. One problem is that any such disclosure could potentially reveal trade secrets or other protected information.[12] Another ongoing issue is that criminal prosecutors are asserting government privilege to prevent disclosure of the algorithms they use to prosecute defendants.[13] Some argue that such algorithms should be treated the same as any other piece of evidence during discovery, allowing subpoenas to take

---

[6] Chander, *supra* note 5, at 1040.

[7] *Id.* at 1038-39; Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1283-84 (2020).

[8] *See* Sharona Hoffman & Andy Podgurski, *Artificial Intelligence and Discrimination in Health Care*, 19 YALE J. HEALTH POL'Y L. & ETHICS 1, 1 (2020).

[9] *See* Theodoros Evgenio et al., *What Happens When AI is Used to Set Grades?*, HARV. BUS. REV. (Aug. 13, 2020), https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades [https://perma.cc/J2HF-RN67].

[10] *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 8:50 PM), https://www.reuters.com/article/idUSL2N1VB1FQ/ [https://perma.cc/XFP4-SYKB].

[11] Chander, *supra* note 5, at 1041; Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CALIF. L. REV. 1539, 1578 (2022); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 681 (2017).

[12] Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1376 (2018); Chander, *supra* note 5, at 1040.

[13] *See generally* Rebecca Wexler, *Ignorance of the Rules of Omission: An Essay on Privilege Law*, 76 VAND. L. REV. 1609 (2023).

effect to reveal information about a protected algorithm.[14] At the same time, the Supreme Court has recognized that disclosure to others of trade secrets given to the government requires compensation under the Takings Clause.[15] Applied here, algorithms could be considered trade secrets, and the government may have to compensate the owners of that intellectual property each time they turn over an algorithm in response to a subpoena. These realities have led Pasquale to argue for only a qualified transparency that acknowledges the limits imposed by intellectual property law.[16]

III. DISCLOSURES ABOUT DATA

In addition to knowing the variables comprising an algorithm and the weights given to those variables, understanding the scope of the data on which an AI model was trained can play a critical role in interpreting its results and knowing the circumstances under which it may be properly used.[17] Simply put, all AI is a form of predictive analytics that identifies patterns in existing data and uses those patterns to generate responses to prompts given to it. As a result, every AI system is necessarily a reflection of the data on which it is trained.

A. Biased Training Data

The most mature part of the commentary on the importance of understanding training data is the literature examining how algorithms often replicate the biases that exist in the data on which they are trained. Sandra Mayson's seminal article, *Bias In, Bias Out*, exemplifies scholarship in this area.[18] Tackling one developing context in which algorithms are increasingly employed, it examines how racial bias in the criminal justice system produces biased risk assessment predictive algorithms.[19] The ways that biased data can lead algorithms to produce biased predictions also led Amazon to abandon its AI-based recruiting tool after the model reproduced the gender biases embodied in the company's past hiring decisions.[20]

B. Scope of the Training Data

Beyond bias, the scope of the data on which an AI model was trained plays a critical role in understanding the circumstances under which it is most likely to produce useful predictions. To use a simple example, ChatGPT-4 was initially trained on data through September 2021[21] and has subsequently been updated on

---

[14] Wexler, *Life, Liberty, and Trade Secrets*, *supra* note 12, at 1376.

[15] Ruckelshaus v. Monsanto Co., 467 U.S. 986, 1003-04 (1984).

[16] Pasquale, *supra* note 3, at 163-65.

[17] Chander, *supra* note 5, at 1039, 1040.

[18] Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019).

[19] *Id.* at 2224.

[20] *See* Dastin, *Amazon Scraps Secret AI Recruiting Tool*, *supra* note 10.

[21] *GPT-4*, OPENAI (Mar. 14, 2023), https://openai.com/research/gpt-4 [https://perma.cc/ 6TG3-HAXS].

data through April 2023.[22] As a result, any queries it receives about factual events taking place after those deadlines are necessarily extrapolations that must inevitably be hallucinations. Similarly, the fact that ChatGPT-2 was trained on Reddit data and that ChatGPT-3 was trained on Wikipedia data makes it likely that those models will reflect any flaws and limitations contained in the data that embody those types of communications.

Recent developments in weather forecasting further illustrate the reality that an algorithm's capability is necessarily constrained by the scope of the data on which it is trained. The traditional approach, which relies on numerical equations that model the physical principles that determine the weather, requires a supercomputer and can take hours.[23] Multiple companies have begun using a new approach based on AI, which utilizes statistical patterns in the data.[24] Early studies indicate that the resulting model produces more accurate results, runs 1,000 to 10,000 times faster, and requires far less computing power than the conventional approach.[25] Interestingly, the AI-based model has also proven more effective in predicting three types of severe weather events, specifically the paths taken by tropical cyclones, extreme heat, and extreme cold.[26] At the same time, concerns remain that the limited amount of training data available may leave these models ill-equipped to forecast rarer extreme weather events.[27] Underrepresentation of these phenomena in the training data may prevent the model from predicting the ways that different factors interact under those conditions.[28] Until these data deficiencies are addressed, even proponents of these models recognize that they are best regarded as complements to, rather than replacements for, conventional forecasting techniques.[29]

The solution preferred by many data scientists to resolve flaws in an algorithm is to throw more data at it. The problem with this solution is that if the scope of the additional data is the same as the original data, adding more of it is unlikely to broaden the range of circumstances under which the algorithm can be expected to yield accurate predictions. Although the discussion of ChatGPT above focuses on temporal scope of the training data, other dimensions of data quality may play important roles as well. In addition to *volume*, the literature commonly discusses

---

[22] Hayden Field, *Microsoft-Backed OpenAI Announced GPT-4 Turbo, Its Most Powerful AI Yet*, CNBC (Nov. 6, 2023, 1:15 PM), https://www.cnbc.com/2023/11/06/openai-announces-more-powerful-gpt-4-turbo-and-cuts-prices.html [https://perma.cc/KU5K-2S5A].

[23] Carissa Wong, *DeepMind AI Accurately Forecasts Weather—On a Desktop Computer*, NATURE (Nov. 14, 2023), https://www.nature.com/articles/d41586-023-03552-y [https://perma.cc/VM9Z-5W4G].

[24] *Id.*

[25] *Id.*

[26] Remi Lam et al., *Learning Skillful Medium-Range Global Weather Forecasting*, 382 SCI. 1416, 1419-20 (2023).

[27] Imme Ebert-Uphoff & Kyle Hilburn, *The Outlook for AI Weather Prediction*, 619 NATURE 473, 474 (2023).

[28] *Id.* at 474; Kroll et al., *supra* note 11.

[29] *See* Lam et al., *supra* note 25, at 1421.

three other "V's" of data quality, including *variety*, *velocity*, and *veracity*.[30] Other discussions analyze data quality in terms of additional dimensions such as accuracy, completeness, consistency, and timeliness.[31] Although data scientists are developing ways to compensate for quality problems,[32] all of these solutions are costly and imperfect.

Even when an algorithm is applied to a problem that falls within the scope of the data on which it was trained, it may still fail to generate accurate predictions if the environment has undergone significant structural changes since the training data was collected. Such a deviation from past patterns is said to have caused the 1998 collapse of Long-Term Capital Management ("LTCM"), an algorithmic hedge fund co-founded by Nobel Laureates Robert Merton and Robert Scholes, among others, that performed so well that it became largest in the world before Russia's default on its debt presented the algorithm with circumstances that it had never previously encountered.[33] Some attribute the multi-billion dollar failure of Zillow's algorithmically driven iBuying platform to the unanticipated changes caused by the COVID-19 pandemic that caused the real estate market to behave in ways different from the past.[34]

These dynamics further reveal the shortcomings of regarding simply throwing more data at the model as a panacea. If the new data possess the same limitations in scope as the existing data, adding them will not broaden the circumstances under which the algorithm is likely to perform well. This has led to the recommendation that algorithms reveal the source of the data on which they were trained.[35] Some have extended this principle to AI,[36] although some have challenged such a requirement as impractical.[37]

---

[30] Michal S. Gal & Nicolas Petit, *Radical Restorative Remedies for Digital Markets*, 36 BERKELEY TECH. L.J. 617, 638 (2021).

[31] *See, e.g.*, Abdulaziz Aldoseri et al., *Re-thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges*, APPLIED SCI., June 13, 2023, at 1, 6. One discussion went so far as to identify 42 "V's." Muhammad Mashab Farooqi et al., *Big Data in Healthcare: A Survey*, *in* APPLICATIONS OF INTELLIGENT TECHNOLOGIES IN HEALTHCARE 143, 14445 (Fazlullah Khan et al. eds., 2019).

[32] *See, e.g.*, Pierre-Alexandre Mattei & Jes Frellsen, *MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets*, PROC. 36TH INT'L CONF. ON MACH. LEARNING 4413 (2019); *see also* David Williams et al., *On Classification with Incomplete Data*, 29 IEEE TRANSACTIONS ON PATTERN ANALYSIS & MACH. INTEL. 427 (2007).

[33] *See* René Stulz, *Why Risk Management Is Not Rocket Science*, FIN. TIMES, June 27, 2000, Special Supp., at 74.

[34] *See* Alix Langone, *What Happened at Zillow? How a Prized Real Estate Site Lost at iBuying*, CNET (Nov. 18, 2021, 10:52 AM), https://www.cnet.com/personal-finance/mortgages/what-happened-at-zillow-how-a-prized-real-estate-site-lost-at-ibuying/ [https://perma.cc/E2DF-3F3N].

[35] *See, e.g.*, Chander, *supra* note 5, at 1040. *See generally* Zarsky, *supra* note 3.

[36] *See, e.g.*, Fariha Tasmin Jaigirdar et al., *What Information Is Required for Explainable AI?: A Provenance-Based Research Agenda and Future Challenges*, PROC. 2020 IEEE 6TH INT'L CONF. ON COLLABORATION & INTERNET COMPUTING (CIC 2020) 177, 180 (2020).

[37] *See, e.g.*, Matthew Elmore, *The Hidden Costs of ChatGPT: A Call for Greater Transparency*, 23 AM. J. BIOETHICS 47, 47 (2023).

Knowing the provenance of training data may play a particularly important role when data or algorithms are repurposed from one function to another.[38] Understanding training data can show where gaps exist that will eventually lead to unexpected outcomes.[39] These considerations have made it increasingly apparent that data disclosure represents an essential component of algorithmic transparency.

C. Constraints on Access to Training Data

Calls for data disclosure must confront the fact that the use of training data is protected by a number of legal regimes. Specifically, data disclosure implicates copyright, privacy, and cybersecurity law.[40]

1. Copyright

A key issue is whether access to copyrighted works as training data represents a violation of copyright law.[41] The question whether the inclusion of copyrighted works in training datasets constitutes fair use remains an open question,[42] although the recent lawsuit brought by the *New York Times* against Microsoft and Open AI may provide more clarity.[43] Recent Supreme Court decisions have found uses of copyrighted materials to be fair use when they add "something new, with a further purpose or different character," such as when they lead to new products.[44] The additional functionality provided by AI may well suffice to meet this criterion.[45]

---

[38] Karl Werder et al., *Establishing Data Provenance for Artificial Intelligence Systems*, 13 ACM TRANSACTIONS ON MGMT. INFO. SYS. art. 22 (2022).

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

[40] The discussion in this section is adapted from Christopher S. Yoo, *Generative AI's Potential Impact on Online Competition*, NETWORK L. REV. (Feb. 13, 2024), https://www.networklawreview.org/yoo-generative-ai/ [https://perma.cc/KP6B-KKNQ].

[41] *Artificial Intelligence and Intellectual Property: Part I — Interoperability of AI and Copyright Law: Hearing Before the Subcomm. on Cts., Intell. Prop., and the Internet of the H. Comm. on the Judiciary*, 118th Cong. 12 (2023) (testimony of Christopher Callison-Burch, Assoc. Professor of Comput. & Info. Sci., Univ. of Pa.), *available at* https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf [https://perma.cc/56YN-5PE8].

[42] Scholars generally favor treating the use of copyrighted works to train GenAI as fair use but recognize that the issue remains unresolved and acknowledge the existence of substantial arguments to the contrary. *See, e.g.*, Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 763-76 (2020); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 301 (2023); Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 68-79 (2017).

[43] Alexandra Brunell, *New York Times Sues Microsoft and OpenAI, Alleging Copyright Infringement*, WALL ST. J. (Dec. 27, 2023, 8:24 AM), https://www.wsj.com/tech/ai/new-york-times-sues-microsoft-and-openai-alleging-copyright-infringement-fd85e1c4 [https://perma.cc/D9PA-DHRG].

[44] Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1202-03 (2021).

[45] Copyright also raises issues of whether returning the verbatim text of a copyrighted work in response to a prompt violates copyright law and whether AI can be considered an author under copyright law. While important, these issues fall beyond questions of the adequacy of data disclosure.

2. Privacy

Privacy law also represents a significant restraint on disclosures about data. In particular, the European Union's General Data Protection Regulation (GDPR) imposes restrictions on the use of personal information that threatens to impede the development of generative AI.[46] As an initial matter, under GDPR, an AI developer must establish that it has a legal basis to process personal data, most likely by arguing that the "processing is necessary for the purposes of the *legitimate interests* pursued by the controller,"[47] unless the training dataset includes "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation," in which case the processor will have to obtain consent from the data subject.[48] GDPR also requires controllers to disclose a range of information to data subjects and provide them with rights of access, rectification, erasure, objection, and data portability, as well as the right to restrict uses under certain circumstances.[49] In addition, GDPR provides data subjects with "the right not to be subject to a decision based solely on automated processing . . . which produces legal effects concerning him or her or similarly significantly affects him or her."[50] These requirements led *Garante per la protezione dei dati personali* (GDDP), the Italian data protection authority, to ban ChatGPT on March 31, 2023,[51] only to restore it without explanation on April 28, 2023.[52] On January 29, 2024, the Italian authorities again notified OpenAI that it believed OpenAI's video AI, Sora, was violating GDPR and gave it thirty days to respond.[53] On March 8, 2024, GDDP opened a formal investigation into Sora.[54]

---

[46] Josephine Wolff et al., *Lessons from GDPR for AI Policymaking*, 27 VA. J.L. & TECH., No. 4, at. 1, 20 (2024).

[47] Regulation 2016/679, arts. 6(1)(4), 2016 O.J. (L 119) 1, 36 (EU). GDPR defines "controller" as the actor that "actor determining the purposes and means of the processing of personal data." *Id.* art. 4(7).

[48] *Id.* art. 9.

[49] *Id.* arts. 13-20.

[50] *Id.* art. 22.

[51] Ashley Belanger, *ChatGPT Data Leak Has Italian Lawmakers Scrambling to Regulate Data Collection*, ARS TECHNICA (Mar. 31, 2023, 2:09 PM), https://arstechnica.com/tech-policy/2023/03/chatgpt-banned-in-italy-over-data-privacy-age-verification-concerns/ [https://perma.cc/R6RA-W8RH].

[52] Kelvin Chan, *OpenAI: ChatGPT Back iIn Italy After Meeting Watchdog Demands*, ASSOCIATED PRESS (Apr. 28, 2023, 2:46 PM), https://apnews.com/article/chatgpt-openai-data-privacy-italy-b9ab3d12f2b2cfe493237fd2b9675e21 [https://perma.cc/ST3U-F7HT].

[53] Kelvin Chan, *ChatGPT Violated European Privacy Laws, Italy Tells Chatbot Maker OpenAI*, ASSOCIATED PRESS (Jan. 30, 2024, 12:09 PM), https://apnews.com/article/openai-chatgpt-data-privacy-italy-a6ff88b53ae611ca4dee917e872ac278 [https://perma.cc/4R4W-8J2R].

[54] Press Release, *Garante per la Protezione dei Dati Personali* (Italian Data Protection Authority), Artificial intelligence: the Italian Data Protection Authority opens an investigation into OpenAI's "Sora" (Mar. 8, 2024), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9991867#english [https://perma.cc/B22S-Z8B9].

Data protection authorities in other countries have reportedly also initiated inquiries of their own.[55]

In addition, scholars are exploring whether certain prompts can cause generative AI systems to leak details of their training datasets in ways that can violate privacy law.[56] Failure to address these privacy concerns could constitute a significant obstacle to the effective deployment of generative AI.

3. Cybersecurity

Generative AI systems must also comply with the laws governing online security. Most notably, the U.S. Computer Fraud and Abuse Act (CFAA) subjects anyone who exceeds their authorized access to a computer to criminal and civil liability.[57] One concern is that some websites make their content available to the public subject to conditions in their terms of service prohibiting wholesale scraping of their data. A recent U.S. Supreme Court decision held that a police officer's breach of such a provision contained in the department's policy did not violate the CFAA while dropping a footnote reserving the question of whether limits in contracts and policies could support CFAA liability.[58] A subsequent Ninth Circuit decision upheld a preliminary injunction supporting a data analytics company's declaratory judgment action that a cease-and-desist letter was insufficient to support a CFAA claim.[59] Although both precedents suggest that the collection of public data that violates terms of service does not violate the CFAA, both stop short of resolving the issue,[60] which potentially places a cloud over any generative AI system trained on data collected in this manner.

* * *

Understanding the training data used to train AI models can yield its own set of challenges. It is no secret that models are trained on vast amounts of data, making it almost impossible for an individual to evaluate the quality of the data without the use of algorithms and other tools to help sift through the unintelligible dataset.[61] Developing a way to understand the scope and quality of the data on which a generative AI model was trained can provide greater insight into the outputs of that model.

---

[55] Chan, *supra* note 53.

[56] Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, ARXIV 1 (Jan. 30, 2023), https://arxiv.org/abs/2301.13188 [https://perma.cc/GGK2-7PRK].

[57] 18 U.S.C. § 1030(a)(1).

[58] Van Buren v. United States, 141 S. Ct. 1648, 1659 n.8, 1660-62 (2021).

[59] hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1195-1201 (9th Cir. 2022).

[60] NAT'L ACADS. OF SCI., ENG'G, & MED., SOCIAL MEDIA AND ADOLESCENT HEALTH 205-06 (Sandro Galeo, Gillian J. Buckley & Alexis Wojtowicz eds., 2023).

[61] Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 195-96 (2017).

IV. DISCLOSURES ABOUT PRE-RELEASE TESTING

Although disclosures about the code comprising algorithms and the data on which they were trained play important roles in helping users understand AI systems, these measures are not sufficient to make generative AI systems accountable.[62] Algorithmic transparency also depends on disclosures about the testing regimes through which an algorithm is validated.

A. The Difficulty in Specifying Objects and Solutions

No testing protocol is perfect, which means that each one necessarily has its strengths and weaknesses. This implies that the details of a testing methodology can play a key role in understanding how well an algorithm is likely to perform and the circumstances under which its predictions are likely to be more uncertain.

Pre-release testing can take many forms, and understanding the components of a particular testing regime can reveal the types of situations that the testing is best suited to assess. For example, a "Standard for Assumptions in Safety-Related Models for Automated Driving Systems," developed by the Institute of Electrical and Electronics Engineers ("IEEE") for autonomous vehicles, employs a variety of methods, including focusing on design processes, compliance with reference architectures, formal methods, robustness analysis, simulation testing, closed course testing, and public road testing.[63] In addition, rather than creating a single analytical framework to cover all aspects of autonomous vehicle safety, the IEEE standard focuses on seven scenarios that the standard developers identified as the most important.[64] Knowing the nature of these scenarios and how they are defined plays a critical role in helping people understand what passing a testing regime means and what aspects it actually validates.

In addition to framing potential strengths and weaknesses, good testing regimes are designed to cover the full range of possible situations the product or service being evaluated is likely to encounter. To cite an example from the non-AI world, seatbelts that had previously passed an automotive crash test later failed when the position of the test weight was shifted in a way that changed the angle of the stress that it placed on the anchors that fastened the seatbelt to the floor.[65] To the extent that these alternative geometries were representative of the circumstances that

---

[62] Cynthia Dwork & Deirdre K. Mulligan, Response, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 37 (2013) ("Exposing the datasets and algorithms of big data analysis to scrutiny—transparency solutions—may improve individual comprehension, but given the independent (sometimes intended) complexity of algorithms, it is unreasonable to expect transparency alone to root out bias.").

[63] IEEE STDS. ASS'N, IEEE STANDARD FOR ASSUMPTIONS IN SAFETY-RELATED MODELS FOR AUTOMATED DRIVING SYSTEMS §§ 6.1-6.6, at 46-48 (2022), https://ieeexplore.ieee.org/document/9761121 [https://perma.cc/NLH7-6U3L].

[64] *Id.* § 4.23., at 26-38.

[65] Kurt D. Weiss, *Failure Mode Testing of Seat Belts*, PLAINTIFF MAG., Jan. 2008, at 1, 6 (noting that the relevant seatbelt standard "unfortunately does not specify that seat belt assemblies be tested in conditions similar to when they are worn" and that "[o]nly through using realistic anchor geometries was the failure mode exposed").

seatbelts are likely to confront in the real world, the first test was not sufficiently robust to evaluate the range of circumstances reasonably likely to occur. This effect is reminiscent of the well-known AI problem of overfitting, in which machine learning tunes an algorithm so closely to the training data that it performs poorly when applied to other data.

Simply put, any program of algorithmic transparency should include disclosures about the way that the system was tested. Such information plays a critical role in promoting an understanding of the circumstances under which an algorithm is likely to perform well.

B. Optimization and Gaming

Testing regimes are also susceptible to strategic behavior known variously as *specification gaming* and *reward hacking*, which occurs when an AI system satisfies the objective given to it in a way that is not consistent with the outcome intended by the designer. This is similar to firms' efforts to artificially promote their search rankings through a method known as search engine optimization ("SEO"), in which website owners attempt to promote their ranking in the results generated by search engines not by improving their product but rather by making changes designed to cater to the selection criteria that the search engine values the most.[66] Real-world experience with SEO underscores the potential tension between transparency and testing, as greater disclosure regarding algorithms opens the door to actors that would engineer their offerings to inflate their search rankings artificially.[67] This dynamic is reflected in "Goodhart's Law,"[68] which is typically quoted as stating that "when a measure becomes a target, it ceases becoming a good measure."[69]

Examples of how AI systems have found ways that comply with the strict letter of their reward criteria while deviating from the designers' intentions are legion.[70]

---

[66] *See, e.g.*, Chander, *supra* note 5, at 1080.

[67] Rachel Pollack Ichou, *Opening the Black Box: In Search of Algorithmic Transparency* 14 (paper presented at the GigaNet 11th Annual Symposium, Dec. 5, 2016), *available at* https://ssrn.com/abstract=3837723 [https://perma.cc/F66N-5JQ3]); Marissa Mayer, *Do Not Neutralise the Web's Endless Search*, FIN. TIMES (July 14, 2010), https://www.ft.com/content/0458b1a4-8f78-11df-8df0-00144feab49a [https://perma.cc/EU6X-H47E].

[68] For the initial statement, see Charles E. Goodhart, *Problems of Monetary Management: The U.K. Experience*, *in* 1 PAPERS IN MONETARY ECONOMICS 91, 116 (1975) (observing that "any observed statistical regularity will begin to collapse once pressure is placed upon it for control purposes"). For a discussion of the history of Goodhart's Law, see K. Alec Chrystal & Paul D. Mizen, *Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy*, *in* 1 CENTRAL BANKING, MONETARY THEORY AND PRACTICE: ESSAYS IN HONOR OF CHARLES GOODHART 221 (Paul Mizen ed., 2003).

[69] Marilyn Strathern, *"Improving Ratings": Audit in the British University System*, 5 EUR. REV. 305, 308 (1997).

[70] *See, e.g.*, Victoria Krakovna et al., *Specification Gaming: the Flip Side of AI Ingenuity*, GOOGLE DEEPMIND (Apr. 21, 2020) https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ [https://perma.cc/ZM9A-5ZF2]; Alex Irpan, *Deep Reinforcement Learning Doesn't Work Yet*, SORTA INSIGHTFUL (Feb. 14, 2018), https://www.alexirpan.com/2018/

Sometimes, these deviations result from bad objective specification. For example, a pancake-flipping bot defined success as the length of time that a pancake could avoid hitting the floor.[71] Although designers intended this to occur through repeated normal flips, the bot optimized its objective criteria by flinging the pancake as high in the air as it could.[72] The problem was that the criterion that the AI was asked to maximize made an imperfect fit with the behavior that the designers sought to train the AI to perform.

In other cases, the problem arises from imperfect specification of the ways an AI system could solve a problem. For example, a neural network known as CycleGAN that was tasked with developing an algorithm that could turn aerial images into street maps and then back into aerial images did so by simply storing a copy of the original aerial image in the code for the street map.[73] In another case, a bot that was asked to play Tetris as long as possible simply put the game on pause.[74] In this case, the problem lay not in the way the designers defined the objective but rather in the failure to place sufficient limits on the ways that the AI could achieve those objectives.

Test criteria can raise ethical concerns as well. A well-known example is the image of Lena Forsén cropped from the centerfold of a 1972 issue of *Playboy* magazine that became a standard test image for digital image processing only to later confront concerns that using such an image unnecessarily alienated women in a male-dominated profession.[75] Testing regimes can thus raise social issues completely unrelated to algorithmic performance.

* * *

Understanding an AI system's likely behavior thus depends on understanding the regime used to test it as much as it does on disclosure of its parameters and the data on which it was trained. Information about the testing regime provides critical details about the types of methods used and the range of circumstances under which it was tested. It also allows users to assess how vulnerable the system is to strategies to yield results that benefit certain actors without promoting the system's overall

---

02/14/rl-hard.html [https://perma.cc/BJ4G-PJMD]; Joel Lehman et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, 26 ARTIFICIAL LIFE 274 (2020).

[71] Robert Miles, *9 Examples of Specification Gaming*, YOUTUBE (Apr. 29, 2020), https://www.youtube.com/watch?v=nKJlF-olKmg [https://perma.cc/JH7P-ZUCA].

[72] *Id.*

[73] Casey Chu et al., *CycleGAN, a Master of Steganography*, ARXIV (Dec. 16, 2017), https://arxiv.org/pdf/1712.02950.pdf [https://perma.cc/3F34-48N7]; David Coldewey, *This Clever AI Hid Data from Its Creators to Cheat at Its Appointed Task*, TECHCRUNCH (Dec. 31, 2018, 6:14 PM), https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/ [https://perma.cc/PW58-CVWR].

[74] Tom Murphy VII, *The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel . . . After That It Gets a Little Tricky.* (Apr. 1, 2013), http://www.cs.cmu.edu/~tom7/mario/mario.pdf [https://perma.cc/4JGT-C5JT].

[75] Jamie Hutchinson, *Culture, Communication, and an Information Age Madonna*, IEEE PRO. COMMC'N SOC'Y NEWSLETTER, May/June 2001, at 1, 5-6.

objectives through opportunistic techniques such as specification gaming and reward hacking.

V. ONGOING POST-RELEASE EVALUATION

The fact that complex systems tend to exhibit emergent behavior that is difficult to predict makes ex ante disclosure of algorithms, data, and testing regimes unlikely to identify all potential problems associated with those systems.[76] These issues apply with even greater force to AI.[77] The difficulty in anticipating these emergent behaviors necessarily requires that any AI standard include some form of ex post auditing and testing.[78] The fact that AI algorithms are constantly learning makes the need for ongoing evaluation all the more acute. Moreover, they can exhibit problematic behaviors that can be detected only after they have been deployed in the real world at scale.

A. Identifying Proxy Discrimination

One area that requires ex post evaluation is algorithmic discrimination. As noted earlier, systems that are prohibited from basing decisions on characteristics such as race, gender, religion, and other similar factors may nonetheless rely on some combination of proxies that replicate those prohibited characteristics.[79] Unless those evaluating an algorithm understand the correlations among all of the possible variables with those variables whose use is prohibited, identification of such proxy discrimination will necessarily involve an examination of the algorithm's outputs.[80]

B. Multiple Agents

Emergent behavior often arises from multiple decisions that are individually rational but interact in ways that cannot be predicted by examining each decision in isolation. The rarity of such incidents has led them to be called *black swan events*, invoking the discovery of black swans in Australia after Europeans long believed that such animals did not exist.[81]

One classic example of this phenomenon from outside the world of AI is the flash crash of May 6, 2010, when an error by a trader triggered a cascade of program

---

[76] For a collection of examples from computer science, see Jeffrey C. Mogul, *Emergent (Mis)behavior vs. Complex Software Systems*, ACM SIGOPS OPERATING SYS. REV., Oct. 2006, at 293, 294-96.

[77] *See* Peter Stone & Manuela Veloso, *Task Decomposition, Dynamic Role Assignment, and Low-Bandwidth Communication for Real-Time Strategic Teamwork*, 110 A.I. 241, 262-71 (1999); Alexander U. Berezhnoy, *Emergent Behavior in Multiagent Systems*, PROC. 3D ANN. WINONA COMPUT. SCI. UNDERGRADUATE RSCH. SYMP. 50, 50 (2003); Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. (forthcoming 2024) (manuscript at 11-21).

[78] EU AI Act Proposal, *supra* note 4, recital 69, arts. 5.3, 7b(b).

[79] *See supra* note 7 and accompanying text.

[80] *See* Chander, *supra* note 5, at 1039-40.

[81] NASSIM NICHOLAS TALEB, FOOLED BY RANDOMNESS: THE HIDDEN ROLE OF CHANCE IN LIFE AND IN THE MARKETS 119-20 (1st ed. 2001).

trades that caused market indices to plummet 9-10% in four-and-a-half minutes only to recover over the next fifteen minutes.[82] During this interval, individual securities traded at prices ranging from as low as a penny and as high as $100,000 per share.[83] At its lowest point, the Dow Jones Industrial Average dropped nearly $1 trillion, one of its largest intraday losses in history.[84] The flash crash was caused by the interaction of the decisions of multiple agents in ways that no one could have anticipated simply by looking at their own planned behavior. Interactions among different actors also reportedly contributed to the collapse of LTCM discussed above.[85]

Unanticipated outputs caused by interaction among multiple agents is not limited to finance.[86] For example, a field experiment by Anja Lambrecht and Catherine Tucker designed to present advertisements promoting job opportunities and training in STEM-related fields equally to women and men found that the algorithm still disproportionately showed the advertisements to men not because of any underlying bias in the data but rather because women were a more highly prized advertising demographic.[87] This caused Tucker and Lambrecht's ads to be outbid by other ads targeted at women.[88] This effect has more sustained implications than the short-term distortions associated with flash crashes.

The real possibility that the behavior of multiple agents can interact in unpredictable ways underscores the need to supplement ex ante disclosure with ex post evaluation. Scholars are now creating models to study the circumstances under which flash crashes can occur for AI.[89] The fact that small perturbations can cause swarming effects suggests that AI can give rise to the type of unanticipated outcomes associated with flash crashes.[90] The unpredictability of such results means that they can only be observed through after-the-fact review once the system has been deployed at scale in the real world.

---

[82] U.S. COMMODITY FUTURES TRADING COMM'N & U.S. SEC. & EXCH. COMM'N, FINDINGS REGARDING THE MARKET EVENTS OF MAY 6, 2010: REPORT OF THE STAFFS OF THE CFTC AND SEC TO THE JOINT ADVISORY COMMITTEE ON EMERGING REGULATORY ISSUES 2-3 (Sept. 30, 2010), https://www.sec.gov/files/marketevents-report.pdf [https://perma.cc/6WEG-EE73].

[83] *Id* at 9.

[84] Christan Borch, *High-Frequency Trading, Algorithmic Finance and the Flash Crash: Reflections on Eventalization*, 45 ECON. & SOC'Y 350, 351 (2016).

[85] *See* TALEB, *supra* note 81, at 242; Stulz, *supra* note 32.

[86] *See* NASSIM NICHOLAS TALEB, THE BLACK SWAN: THE IMPACT OF THE HIGHLY IMPROBABLE (2007) (applying the logic of black swan effects beyond finance).

[87] Anja Lambrecht & Catherine Tucker, *Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads*, 65 MGMT. SCI. 2966, 2976 (2019).

[88] *Id*.

[89] *See, e.g.*, Lorenzo Barberis Canonico & Nathan McNeese, *Flash Crashes in Multi-Agent Systems Using Minority Games and Reinforcement Learning to Test AI Safety*, PROC. 2019 WINTER SIMULATION CONF. (WSC) 193, 193.

[90] *Id*. at 195.

C. Adverse Environments

The foregoing discussion on multiple agents reveals that AI systems can yield unpredictable outcomes even when users take actions that are consistent with the goals of the model. The possibility of unpredictable outcomes becomes even more pronounced in hostile environments populated by those who are not necessarily committed to the best interests of the system.

A prime example of how hostile actors can cause an AI system to go astray is the Microsoft chatbot known as Tay, which devolved into a cesspool of racism, sexism, and antisemitism over the course of sixteen hours after trolls discovered its tendency to learn from and parrot back the content in its Twitter feed.[91] Studies suggest that such toxicity can also emerge in generative AI platforms such as ChatGPT due to bad actors.[92] The real possibility of adverse environments populated by hostile actors provides yet another situation that might only become manifest after the fact.

D. Hallucinations

Even when AI acts in a single-agent environment, it can produce unsatisfactory outputs. Foundation models are quite susceptible to providing information that may misinform its users.[93] Large language models build responses to prompts one word at a time based on the patterns in their training data, a process that has been called "stochastic parroting" because responses are constructed based on probabilities without any reference to meaning.[94] This parroting technique makes it difficult to distinguish between factually correct and incorrect outputs without prior knowledge.

ChatGPT's construction of responses word by word based on probable correlations inferred from the training data sometimes leads it to provide erroneous information.[95] Because of this, ChatGPT provides a footnote at the bottom of its chat feature that reads, "ChatGPT can make mistakes, consider checking important information."[96] Scholars have nonetheless been caught publishing works and

---

[91] Vinayak Mathur et al., *Intelligence Analysis of Tay Twitter Bot*, PROC. 2D INT'L CONF. ON CONTEMP. COMPUTING & INFORMATICS (IC3I) 231, 231 (S.K. Niranjan & V.N. Manjunatha Aradhya eds., 2016).

[92] Ameet Deshpande et al., *Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models*, ARXIV (Apr. 11, 2023), https://arxiv.org/abs/2304.05335 [https://perma.cc/7YQR-Q6CH].

[93] Vipula Rawte et al., *A Survey of Hallucination in Large Foundation Models*, ARXIV (Sept. 12, 2023), https://arxiv.org/abs/2309.05922 [https://perma.cc/3UCL-CHQJ].

[94] Emily Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, PROC. 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610, 616-17 (2021).

[95] *See* ChatGPT 3.5, https://chat.openai.com/share/0e9caec7-b52d-425c-8f6f-81e59d2b88fa [https://perma.cc/BME2-NNN2] (last visited Mar. 3, 2024) (reporting a sequence in which ChatGPT misreports the author's clerkship history).

[96] *ChatGPT 3.5*, https://chat.openai.com/ [https://perma.cc/F2BE-L8WB] (last visited Mar. 22, 2024).

lawyers have been caught submitting briefs generated in whole or in part by foundation models that contained fictitious information and citations.[97]

Sometimes the likelihood that a ChatGPT response is a hallucination is easy to predict. For example, as noted earlier, GPT-4 was initially trained on data through September 2021.[98] This meant that any factual inquiries about facts occurring after that date were necessarily hallucinations. The recent update to include data through April 2023[99] simply changed the date after which hallucinations would occur. For other types of facts, hallucinations are much less predictable. As one Google employee noted, the fact that a query to the company's large language model known as Bard returned a nonexistent accomplishment by the James Webb Space Telescope underscores how generative AI can hallucinate and shows why generative AI must be put through a rigorous ex post evaluation process.[100]

* * *

The tendency for complex systems like AI to exhibit emergent behavior makes clear the need for ongoing evaluation of the performance of AI systems. That said, the details about how to conduct this ex post testing largely remain to be defined. Future work will have to specify exactly what ongoing testing should be required of AI.

VI. CONCLUSION

Understanding AI's impact thus requires much more than disclosure of the terms of the underlying algorithm. It also requires disclosure of many aspects of the data on which the algorithm was trained, including its source, scope, quality, and inner correlations. Further, it requires an appreciation for the tests used to validate it, and, most importantly, the reality that complex systems tend to exhibit emergent phenomena that are difficult to anticipate when operating in the real world. Future standards to govern AI must take these different dimensions of AI transparency into account.

---

[97] Mata v. Avianca, Inc., No. 22-cv-1461 (PKC), 2023 WL 4114965, at *15 (S.D.N.Y. June 22, 2023) (opinion and order on sanctions); Paulina Okunyté, *Google Search Exposes Academics Using ChatGPT in Research Papers*, CYBERNEWS (Nov. 15, 2023), https://cybernews.com/news/academic-cheating-chatgpt-openai/ [https://perma.cc/VYF3-LLC6].

[98] *See supra* note 21 and accompanying text.

[99] *See supra* note 22 and accompanying text.

[100] Catherine Thorbecke, *Google Shares Lose $100 Billion After Company's AI Chatbot Makes an Error During Demo*, CNN (Feb. 9, 2023, 9:41 AM), https://www.cnn.com/2023/08/29/ tech/ai-chatbot-hallucinations/index.html [https://perma.cc/2LLS-QVHM].