REPORT

Opening Black Boxes: Addressing Legal Barriers to Public Interest Algorithmic Auditing

BY NANDITA SAMPATH
POLICY ANALYST
OCTOBER 2022

CR Consumer Reports®

Executive Summary

Artificial Intelligence (AI) is being integrated into everyday decision-making in practically every commercial sector in the U.S., from housing to education to the criminal justice system. Landlords have used automated tenant screening reports (which include an algorithmically generated score) to make determinations about potential tenants.[1] The COVID-19 pandemic has led to schools requiring students to download proctoring software to identify cases of cheating during at-home exams.[2] In the criminal justice system, risk assessments have been used to, among other things, quantify a defendant's future risk of misconduct to determine whether they should be incarcerated before their trial.[3] But as AI-enabled decision-making becomes more common, it also has the potential to exacerbate historical societal inequalities if it generates unfair and biased outcomes.

Before we can regulate algorithms effectively, both regulators and the public need to know how they work and arrive at their conclusions and to what extent they perpetuate discrimination and other harms. While federal and state civil rights laws prohibit discrimination based on protected characteristics like race, gender, and skin color in employment, housing, and lending, it can be hard to detect whether certain algorithms lead to discrimination at all. With many algorithms, it can be difficult to determine how they arrive at their final decisions, even for the engineers who design them.[4] While this paper focuses on identifying discrimination, some companies make unsubstantiated claims about their algorithms, promoting both high accuracy rates and that their algorithms are capable of making certain determinations without external validation.[5] Furthermore, there are few transparency requirements for businesses to disclose how their algorithms work, the types of data they collect, how each data point is factored into the final decisions, and accuracy or error rates.

Ultimately, our government must be the one to set standards on algorithm testing and auditing, particularly for applications with significant legal effects. However, in the absence of laws that require companies using AI to undergo independent, rigorous third-party audits, public interest researchers can play a vital role in uncovering the harms caused by algorithmic decision-making. This paper will lay out the different types of public interest auditing techniques and then address the legal and practical roadblocks that can impede public interest researchers from performing algorithmic audits. Public interest audits are limited by imperfect access to

---

[1] Kaveh Waddell, "How Tenant Screening Reports Make It Hard for People to Bounce Back From Tough Times," Consumer Reports, March 11, 2021, https://www.consumerreports.org/algorithmic-bias/tenant-screening-reports-make-it-hard-to-bounce-back-from-tough-times-a2331058426.

[2] Drew Harwell, "Cheating-detection companies made millions during the pandemic. Now students are fighting back," The Washington Post, November 12, 2020, https://www.washingtonpost.com/technology/2020/11/12/test-monitoring-student-revolt.

[3] Alex Chohlas-Wood, "Understanding risk assessment instruments in criminal justice," Brookings Institution, June 19, 2020, https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice.

[4] Will Knight, "The Dark Secret at the Heart of AI," MIT Technology Review, April 11, 2017, https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai; Roman V. Yampolskiy, "Unexplainability and Incomprehensibility of Artificial Intelligence," *Journal of Artificial Intelligence and Consciousness* 7, no. 2 (June 20, 2019), https://philarchive.org/archive/YAMUAI.

[5] Arvind Narayanan, "How to recognize AI snake oil," Princeton University, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

algorithms and the underlying data in part because of existing laws designed to limit computer hacking and protect intellectual property. To help remove these obstacles, we recommend policy changes that would balance these legitimate values with the need for research and external accountability. Today, public interest researchers are significantly hindered in performing good faith research to identify sources of algorithmic harm because they are concerned about a potential lawsuit. Policymakers should make targeted changes to the law to address this chilling effect.

Table of Contents

Introduction

Artificial intelligence (AI) and Machine Learning (ML) refer to the use of data to make predictions or classifications about future data points, while an algorithm is simply a set of instructions to make these predictions and classifications. Although there is no consensus over these definitions, both AI and ML generally refer to the types of algorithms used in making these decisions,[6] and sometimes these terms are used interchangeably. In general, though, data is used to train an algorithm so that it can make more accurate decisions, and the algorithm is only as good as the quality of the data it is fed.

As the use of algorithms and AI become more embedded into daily life, the potential for algorithmic harms like discrimination is alarming. There are minimal regulations and industry standards to guide how algorithms are designed and tested, and how to address any negative impacts, and it is often unclear how existing law applies to these new technologies.[7] Because many algorithms are quite complex, it is difficult to regulate them appropriately. However, effective audits by public interest researchers can help both the public and regulators understand how algorithms work and their impact on potential discrimination and other harms.

Mandatory, independent, and standardized third-party audits for companies whose algorithms pose significant legal effects are vital for maintaining our civil rights as more processes that affect our lives become automated. This could be done by either government agencies or private companies that have been accredited through a process specified by government agencies that enforce particular laws. For example, the Department of Housing and Urban Development would need to design what an audit should look like to examine algorithms covered under the Fair Housing Act and would need to accredit private auditing companies to carry out these audits, or perform the audits internally.[8]

However, there is a long way to go before this becomes a reality. The U.S. has not yet passed significant AI legislation at the federal level and lags behind governments like the European Union when it comes to enacting technology regulation; and, furthermore, involved federal agencies would likely be limited by funding and staffing issues in order to carry out audits or create an accreditation process effectively. While the burden in the meantime should not fall entirely on public interest researchers to uncover algorithmic harms, they can play a vital role in identifying bias and calling out companies as we push for more government regulation. And lessons learned from public interest audits can potentially be applied once a regulatory regime is in place.

---

[6] However, AI is more commonly associated with newer types of algorithms such as neural networks that, while having the potential for high accuracy rates when performing difficult tasks (such as visualizing the surroundings of a self-driving car), are also so complicated that even the engineers that design them cannot fully explain how they work. See: Will Knight, “The Dark Secret at the Heart of AI”; Roman V. Yampolskiy, “Unexplainability and Incomprehensibility of Artificial Intelligence.” ML is often used to refer to older, statistical methods like linear regression models and decision trees, for example, that can more easily be interpretable by engineers and statisticians. These are types of models that can make a prediction or classifications about the output of a system given a particular input.

[7] Mark MacCarthy, “AI needs more regulation, not less,” Brookings Institution, March 9, 2020, https://www.brookings.edu/research/ai-needs-more-regulation-not-less.

[8] Specific frameworks for what this could look like are out of scope for this paper.

Unfortunately, there are many roadblocks that prevent public interest researchers from performing algorithmic audits. The same laws that were created to promote science and art, and to protect individuals and companies from hacking, are also hindering researchers in performing meaningful audits, for fear of legal recourse. These laws include the Computer Fraud and Abuse Act, copyright law, and contract law. Our conclusion is that these laws need to be clarified and updated so that public interest researchers can perform good faith audits without being concerned about legal repercussions.

Problem

Algorithms are often used in place of human decision-making, and in some cases they are touted as being more objective and thorough than a human decision-maker.[9] However, an algorithm is only as good as the engineer who designs it and the data it is trained on—human error, including biased data collection methods and the type of algorithm that is chosen by the engineer, can also cause bias. No algorithm will ever be perfect, because a model is a simplified version of real-world events. Most algorithms make mistakes — or are more accurate on certain groups than others[10] — due to these errors during the design process. This can cause real harm when the algorithm is used by a government, school, workplace, or even a landlord.[11]

While there are some laws that prohibit discrimination based on protected characteristics like race, gender, and skin color in employment, housing, and lending, it is often difficult to identify whether models used in these areas actually contribute to unequal outcomes based on these characteristics. Companies are typically not required to disclose how their algorithms work, how they trained them, what issues they identified with their technology, and what steps they took to mitigate harm.[12] Furthermore, people usually do not know how the algorithm works on others, so it could be difficult for them to even identify whether they were discriminated against (for example, a woman who is rejected for a job by a resume-screening algorithm may not know that it allowed a man of similar experience to pass through).

Algorithmic discrimination is not the only harm associated with AI—social media platforms have been accused by critics of optimizing their algorithms for engagement, which leads to the spread of misinformation, propaganda, and harmful targeted advertisements.[13] Many companies

---

[9] Rebecca Heilweil, “Artificial intelligence will help determine if you get your next job,” Vox, December 12, 2019, https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen; Sendhil Mullainathan, “Biased Algorithms Are Easier to Fix Than Biased People,” The New York Times, December 6, 2019, https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html.

[10] The National Institute of Standards and Technology found that certain facial recognition algorithms were more likely to misidentify Asian and African American faces relative to Caucasians. “Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects,” National Institute of Standards and Technology: News, December 2019, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf.

[11] There are entire books written about these issues, such as *Weapons of Math Destruction* by Cathy O'Neil (Crown Publishing Group, 2016) and *Race After Technology* by Ruha Benjamin (Polity, 2019).

[12] Hannah Bloch-Wehba, “Transparency’s AI Problem,” Knight First Amendment Institute at Columbia University, June 17, 2021, https://knightcolumbia.org/content/transparencys-ai-problem.

[13] Filippo Menczer, “How ‘engagement’ makes you vulnerable to manipulation and misinformation on social media,” The Conversation, September 10, 2021, https://theconversation.com/how-engagement-makes-you-vulnerable-to-manipulation-and-misinformation-on-social-m

also promote their AI as being capable of predicting social outcomes or other kinds of "snake oil." In other words, they make claims about their products that are not backed up by science.[14] Many companies tout their "emotion recognition" algorithms, claiming they can identify how someone is feeling based on their face or other physical characteristics; there are concerns that these algorithms could discriminate based on race and have other harmful implications, and there is no evidence that emotion recognition can be done accurately.[15] Algorithmic discrimination can lead to other egregious, distinct harms—consider hospitals using historical data about patients in an algorithm intended to help decide how to triage patients. One paper found that Black patients were assigned lower-risk scores than white patients, even when they were equally sick; the algorithm used data about patients' historical healthcare costs to make decisions, and Black patients were routinely spent less on, which the scientists speculated is due to systemic barriers to healthcare access.[16] Oversights like these are a matter of life or death, and we should expect robust standards for these kinds of algorithms.

Ultimately, AI can exacerbate power imbalances between consumers and companies (endless data collection about a consumer can lead to discriminatory pricing for products, or can be used to nudge a consumer to behave a certain way on a platform). AI companies need to be held accountable for AI-enabled harm, and they should be required to make transparent their accuracy rates and testing procedures, or otherwise change their algorithm design and testing procedures when such harms are identified.

Case for Public Interest Auditing

An algorithmic audit can be instrumental in identifying and mitigating algorithmic harm. An audit can help determine whether an algorithm leads to unequal outcomes or harmful effects. It can also identify in what context an algorithm works well, and when it fails. Ultimately, the purpose of an algorithmic audit is highly dependent on the auditor's goals and the information they have access to in carrying out the audit.

Specifically, we argue that public interest groups, academics, and journalists have a major role to play in identifying algorithmic harms[17] (in the absence of and alongside future government regulation of algorithms) because, unlike private auditing companies hired by the AI companies

---

edia-145375; Steve Dent, "Facebook whistleblower reveals identity, says company chooses 'profits over safety,'" October 4, 2021, https://techcrunch.com/2021/10/04/facebook-whistleblower-reveals-identity-says-company-chooses-profits-over-safety.

[14] Arvind Narayanan, "How to recognize AI snake oil," Princeton University, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

[15] Lisa Feldman Barrett et al., "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," Psychological Science in the Public Interest, July 17, 2019, https://journals.sagepub.com/doi/10.1177/1529100619832930#_i72.

[16] Heidi Ledford, "Millions Affected by Racial Bias in Health-Care Algorithm," *Nature* 574 (October 31, 2019): 608-609, https://media.nature.com/original/magazine-assets/d41586-019-03228-6/d41586-019-03228-6.pdf.

[17] For example, ProPublica was able to look at outputs of the COMPAS algorithm (which claims to predict a criminal defendant's likelihood of becoming a recidivist), to determine that the algorithm often predicted Black individuals to be at a higher risk of recidivism than they actually were, and white individuals were often predicted as less risky than they actually were; Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

themselves, they typically seek to make available to the public useful information about how algorithms work, and to determine whether these algorithms lead to discriminatory or other harmful outcomes. We define a public interest algorithmic audit as investigatory research into an algorithm intended to discover and inform the public about potential harms caused by the algorithm. They can be performed by academics, public interest groups, journalists, or just concerned citizens. However, these investigators need access to adequate information in order to perform effective audits (which they do not always have).

Why Private Audits Are Not Enough

In contrast, private audits can be ineffective without basic auditing requirements and standards.[18] Because the AI company is the one paying the private, third-party auditor (and generally there are no legal requirements for most AI companies to undergo an audit[19]), the AI company can essentially set its own standards for what the audit should entail, which could lead to weak and rather meaningless accountability measures.[20] AI companies can determine what types of audits they want to undergo, what specific algorithms they want to be audited, and how much of their information they want to give to auditors (even under a nondisclosure agreement). Companies can also choose which products to audit, keeping secret the ones that are failing while presenting a good public image. It is also likely that different auditing companies will have wildly different techniques in terms of which issues they search for and how they go about identifying them—Auditor A might obtain a significantly different impact assessment of a company's algorithm than Auditor B.

In the absence of auditing transparency requirements, companies that voluntarily undergo audits by private auditing companies can mischaracterize the results in a way that is misleading to the public. Private auditing companies offer auditing services to AI companies. However, because there are few, if any, legal requirements for a third-party audit,[21] it is not clear that these services will identify or mitigate potential harms.[22] A company could use inadequate private

---

[18] Consider the case of the Arthur Andersen and Enron scandal. The firm Arthur Andersen served as both a consultant and auditor for Enron, which was a conflict of interest, and led to Arthur Andersen being indicted for obstruction of justice after destroying Enron audit information requested by the SEC (which essentially resulted in the downfall of both companies). Ken Brown and Ianthe Jeanne Dugan, "Arthur Andersen's Fall From Grace Is a Sad Tale of Greed and Miscues," The Wall Street Journal, June 7, 2002, https://www.wsj.com/articles/SB1023409436545200.

[19] Proposed legislation such as the Algorithmic Accountability Act (S.3572 and H.R. 6580) and Washington State's SB 5116 (which failed in March 2022) would require auditing, but there are currently no industry-wide or legal standards to determine the kinds of information companies should provide to auditors about their technology in order for an audit to take place, and even what the audit should address. Because AI applications are diverse and varied, these standards need to be nuanced based on the context of the algorithm. One exception is a New York City law that would require a bias audit be conducted on an automated employment decision tool prior to the use of said tool.

[20] Megan Gray, "Understanding and Improving Privacy 'Audits' under FTC Orders," The Center for Internet and Society: Stanford University, April 2018, http://cyberlaw.stanford.edu/files/blogs/white%20paper%204.18.18.pdf.

[21] The Federal Trade Commission has put out business guidelines for developing and using AI (https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms) that include testing algorithms for bias, making sure decisions are explainable to consumers, and more. While there are not necessarily laws that require testing in a particular way or at all, antidiscrimination law and other laws like Section 5 of the FTC Act could hold companies accountable for failing to identify and mitigate disparate impacts in their algorithms, stated here: https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

[22] Alfred Ng, "Can Auditing Eliminate Bias from Algorithms?" The Markup, February 23, 2021, https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms.

audits to justify its business practices, rather than to address the potential harms caused by them. HireVue, a video software company that claimed to analyze people's faces during the job interview process, obtained the services of O'Neil Risk Consulting & Algorithmic Auditing (ORCAA) after the company had been widely criticized for allegedly being biased and using debunked pseudoscience to score applicants.[23] However, it was audited only for a narrow hiring test rather than its "candidate evaluation process as a whole."[24] HireVue claimed in a press release that the audit was successful, though the audit addressed only a specific issue.[25]

Finally, companies being audited typically are not required to disclose results of these audits to the public, or to address any problems identified in the audit. The lack of transparency or risk mitigation requirements means companies can tout the fact they have undergone an audit (which can make them look more ethical or responsible as a company) without actually meaningfully addressing the issues identified by the audit.

Public interest audits generally lack the monetary incentives of private audits, and are done to uncover new information and identify potential issues that algorithms pose. Journalists and researchers generally play a part in providing the public with information in regards to issues that companies pose to the public, such as corruption, fraud, and more. And journalists and researchers should be given the same opportunities to do the same with AI companies, which in some cases can pose harm to the public or end-users of an AI application.

For example, researchers at New York University conducted a study called the Ad Observatory, where they obtained consent from volunteer Facebook users who gave the researchers access to the ads the users were seeing on their newsfeed. This study gave the researchers insight into how political ads were algorithmically targeted to users, and the collected ads were put into a publicly available database for other researchers and journalists to examine.[26] While Facebook has an advertisement database available to the public that it claims contains all political ads shown to users, the Ad Observatory group found that Facebook routinely misses including political ads in this database[27] and sometimes fails to disclose who pays for some political ads.[28]

It is not always seasoned researchers who can identify problems with algorithms. Twitter users noticed in 2020 that Twitter's image-cropping algorithm, which showed a preview of an image on a user's feed, was perhaps biased toward younger, slimmer, and lighter faces.[29] Due to the

---

[23] The company discontinued the use of "visual analysis" from its job interview assessments in early 2020. https://www.hirevue.com/press-release/hirevue-leads-the-industry-with-commitment-to-transparent-and-ethical-use-of-ai-in-hiring.
[24] Alfred Ng, "Can Auditing Eliminate Bias from Algorithms?"
[25] Id.
[26] Ultimately, Facebook ended up disabling the researchers' accounts, effectively ending the study. Lois Anne DeLong, "Facebook Disables Ad Observatory; Academicians and Journalists Fire Back," NYU Center for Cybersecurity, August 21, 2021, https://cyber.nyu.edu/2021/08/21/facebook-disables-ad-observatory-academicians-and-journalists-fire-back.
[27] Nancy Watzman, "The political ads Facebook won't show you," Cybersecurity for Democracy, Medium Blog, May 12, 2021, https://medium.com/cybersecurity-for-democracy/the-political-ads-facebook-wont-show-you-e0d6181bca25.
[28] Shirin Ghaffary, "People do not trust that Facebook is a healthy ecosystem," Vox, August 6, 2021, https://www.vox.com/recode/22612151/laura-edelson-facebook-nyu-ad-observatory-social-media-researcher.
[29] Alex Hern, "Student proves Twitter algorithm 'bias' toward lighter, slimmer, younger faces," The Guardian, August 10, 2021,

backlash, Twitter ended up giving more information on how it tests for bias in the image-cropping model, and also gave users more control over how their images were cropped before being published.[30]

Clearly, public interest researchers can play a big role in identifying and mitigating harm posed by algorithms. However, the type of audit that can be executed and the extent to which a researcher is able to assess a model is highly dependent on the information they have access to. In the next few sections, we will discuss the different types of algorithmic audits, and the practical and legal limitations of each, and suggest policy recommendations to remove barriers to make it easier for researchers to conduct these audits.

https://www.theguardian.com/technology/2021/aug/10/twitters-image-cropping-algorithm-prefers-younger-slimmer-faces-with-lighter-skin-analysis.

[30] Parag Agrawal and Dantley Davis, "Transparency around image cropping and changes to come," Twitter Blog, October 1, 2020, https://blog.twitter.com/official/en_us/topics/product/2020/transparency-image-cropping.html.

Introduction to Types of Audits

Public interest groups, academics, and journalists have a major role to play in identifying algorithmic harms, but legal and practical roadblocks often prevent public interest researchers from performing effective AI audits. The same laws that were created to promote science and art, and protect individuals and companies from hacking, are unfortunately also hindering researchers from performing meaningful audits, for fear of legal recourse. These include laws like the Computer Fraud and Abuse Act and copyright law, as well as tort and contract law.

Below, we will describe different types of audits, including code audits, crowdsourced audits, scraping, and sock puppet audits, and their practical and legal limitations for auditing algorithms to identify discrimination or other harms. All of the audit practices described are essential to conducting research on algorithmic discrimination and other harms. The auditor often selects the type of audit based on the availability of information about the system, as well as the resources they have to conduct the audit. Typically, researchers are limited in both access to the necessary information and resources to conduct the audit. Because the purpose of audits is to understand how and when a system works as well as when it fails, researchers need input or output data of a system, adequate staff, and powerful computers to conduct an effective audit.

The audit categories below are fairly generalized—there exist auditing practices that combine any or all of the categories and also practices that are perhaps more nuanced than any of the descriptions below. The categories chosen are derived from Christian Sandvig's paper on algorithmic audits, but examples and categories have been changed slightly for the purposes of this paper to help distinguish between some of the legal and practical issues that exist between them.[31]

Each type of audit, when carried out by public interest groups, has both practical and legal limitations. We identify the main limitations posed by each audit—and where clear legal barriers exist, we suggest policy solutions to remove them.

1. Code Audit

Description:

The first type of audit is fairly straightforward: A code audit is when an auditor gains access to a company's source code, which can be the underlying code of any model or algorithm. For example, Twitter made its image-cropping code public after it received backlash about the code's potential biases.[32] The public was able to review the code and test it to identify sources

[31] Christian Sandvig, Kevin Hamilton, Karrie Karaholios, and Cedric Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," International Communication Association, May 22, 2014, http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

[32] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra, "Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency," arXiv, September 9, 2021, https://arxiv.org/pdf/2105.08667.pdf; Twitter research, Image crop analysis code, GitHub, https://github.com/twitter-research/image-crop-analysis.

of bias.[33] Recently, Elon Musk has suggested making Twitter's algorithms open source to increase trust.[34]

In a code audit, a company can provide the auditor with either the entire codebase or the code regarding any potentially concerning aspects of the software or algorithm. In the case of algorithmic auditing, companies may also need to provide the auditor with training data and other relevant information so that the auditor can test out the system in a robust manner; often, auditors need this extra information to gain a full understanding of how the system works under different circumstances.

Practical Limitations:

Even with full access to an algorithm's code, a code audit on its own may not be useful to the auditor. Even to a sophisticated auditor, it may be difficult to look through thousands or millions of lines of code to identify sources of bias or harm.

Access to the code itself may also be insufficient on its own. It is also generally difficult to test an algorithm without using training and sample input data along with the algorithm itself.[35] If a company chooses to disclose its algorithm but not the data it uses, identifying discrimination and other harms may not be possible because the harms may arise only in the context of specific data usage or interaction with a user.[36]

Finally, few companies are incentivized to make their code available for third-party auditing. Many treat their code as a competitive advantage and might worry that even data shared with one external partner could wind up in the hands of a competitor. If the code became widely available, bad faith actors could potentially find loopholes to game. For example, Google guards its search results algorithms closely and constantly adjusts them to combat search engine optimization efforts that could result in less relevant results for users. On the other hand, providing access to the code could reveal instances of bias or discrimination, subjecting the company to public embarrassment or even potential liability.

Legal Limitations:

Companies currently have little to no legal obligation to release their code to auditors or regulators. To the extent that an auditor tries to access or reverse engineer code without the company's permission, they risk violating hacking laws like the Computer Fraud and Abuse Act or the Digital Millennium Copyright Act, which prohibits circumventing technical measures to protect copyrighted material. And as mentioned above, access to an algorithm's underlying code

---

[33] Curt Wagner, "Hackathon Points to More Biases in Twitter Algorithm," PMCA, August 18, 2021, https://www.pcma.org/defcon-hackathon-finds-more-biases-twitter-algorithm.
[34] Twitter, Inc., April 25, 2022, https://www.prnewswire.com/news-releases/elon-musk-to-acquire-twitter-301532245.html.
[35] Amanda Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem," *Washington Law Review* 93, no. 2 (2018): 628, https://robotic.legal/wp-content/uploads/2018/09/SSRN-id3024938.pdf.
[36] For example, identifying the kinds of ads a person sees on Facebook's newsfeed cannot be done with just the algorithm alone—ads are deployed based on a user's interaction with a platform, so a researcher would need access to information such as how the user has interacted with other users or what they have previously clicked on in order to get a better picture of how the algorithm deploys ads for that individual.

is not usually enough—training data and other contextual data is necessary to robustly audit an algorithm.

In addition to copyright protections over the code, training datasets themselves could include copyrighted content like artwork or copyrighted text. Regardless of whether the company had the legal right to use the copyrighted images, researchers attempting to use the training data either to test algorithms or to reverse engineer potentially problematic algorithms could run into the issue of copyright infringement, even if the company willingly made it available.[37]

The potential for exposure to liability for such infringement may disincentivize companies from releasing their datasets in the first place.[38] Also, depending on the originality of the selection and arrangement of the information in the dataset, companies might try to claim copyrightability over the dataset itself, disincentivizing research from other parties—public interest, adversarial, or otherwise—for fear of infringement litigation, regardless of whether or not the use may ultimately be fair.[39]

2. Crowdsourced Audit

Description:

A crowdsourced audit is essentially a survey of users to gather data about their normal interactions with an algorithm or platform (for example, getting users to share all of their queries on a search engine). An auditor can get volunteers to either provide information about their interactions with the algorithm or provide direct access to the auditor (with consent) to view their interactions. For example, Consumer Reports has previously done similar participatory research to identify differences in insurance cost estimators offered to consumers and to identify roadblocks for consumers trying to exercise their rights under the California Consumer Privacy Act.[40]

Practical Limitations:

While crowdsourced audits can be extremely useful in shining a light on companies' practices, they do have some important practical limitations. First, testers will self-select, and may not be representative of the general population, unless researchers make careful choices about which testers to use. For example, volunteer testers may already have strong opinions about the

---

[37] Consider an example where an algorithm was developed to look at images of flowers and classify them by species. If the training set of flower images was scraped from various photography websites that specialized in nature photography, researchers attempting to reverse engineer the algorithm would likely need to obtain flower images from similar websites. This could be a copyright violation if they did not obtain permission from the owners (because this can often be costly). Larger technology companies may have the resources to pay for damages due to copyright violation and could be willing to take the risk of using these images without owner permission, while public interest researchers may not have the same ability to do so.

[38] Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem," 597, footnote 77.

[39] *Feist Publications, Inc., v. Rural Telephone Service Co., Inc.,* 499 U.S. 340 (1991), https://cyber.harvard.edu/people/tfisher/1991%20Feist.pdf.

[40] "How We Rate Health Insurance Plan Tools and Public Price Estimator Tools," Consumer Reports, November 2016, https://article.images.consumerreports.org/prod/content/dam/cro/news_articles/health/PDFs/Consumer_Reports_Health_Insurance_Tool_Ratings_Technical_Report.pdf; Maureen Mahoney, Ginny Fahs, and Don Marti, "The State of Authorized Agent Opt Outs Under the California Consumer Privacy Act," Consumer Reports, February 2021, https://advocacy.consumerreports.org/wp-content/uploads/2021/02/CR_AuthorizedAgentCCPA_022021_VF_.pdf.

product or interact with it in particular ways that can skew a sample, similar to how people with strong opinions are often more likely to respond to surveys or give ratings. To be most helpful, the users sampled must exhibit a variety of attributes in order to properly identify discrimination or other potential harms.

Even with a good sample, it can be difficult for researchers to identify causality between the inputs and outputs of the algorithm; outputs could be the result of any number of factors, including previous interactions between a user and the system (which could affect future interactions, like search engine results or advertisement suggestions), which may not properly be identified to researchers.[41] There can also be self-reporting errors made when users share information with the researcher. The use of sock puppet audits (*see infra*, #4) that afford researchers more control over inputs could solve many of the issues presented by crowdsourced user audits, though they also present different legal and practical challenges.

Legal Limitations:

A company's terms of service agreement could purport to limit users' participation in certain audits. For example, a website could prohibit a researcher performing a crowdsourced audit from using a volunteer's information to log in to collect data, even when the volunteer gives consent to do so, or prohibit individuals from disclosing information about their accounts or user experiences to researchers or the public.[42] Companies have broad discretion in crafting website terms of service, and could potentially try to use contract language to frustrate crowdsourcing. In some cases, courts and regulators have found that contractual provisions that limit the publication of testing results are legally "unconscionable" or contrary to public policy.[43] Furthermore, the Consumer Review Fairness Act prohibits contracts from preventing consumers from giving honest reviews about a product or service.[44] However, sometimes companies may have a legitimate interest in preventing their users from sharing certain data related to their products, especially if sharing could infringe the rights of others. For example, Facebook cited concern about others' privacy when shutting down Cambridge Analytica's access to Facebook's APIs after it had exposed a loophole that allowed Cambridge Analytica to collect data not only of individuals who had taken a particular online quiz but also of their Facebook friends.[45]

There is also some legal uncertainty whether violating terms of service (ToS) agreements constitutes a violation of the Computer Fraud and Abuse Act (CFAA) or other state computer hacking laws. A prosecutor could allege that accessing a computer service in contravention of its stated terms and conditions could constitute illegal hacking. The Supreme Court recently ruled in *Van Buren v. United States* that an individual given access to a database but who

---

[41] Christian Sandvig et al., "Auditing Algorithms," 11.
[42] James Snell, Nicola Menaldo, and Ariel Glickman, "CFAA Decision May Raise Bar On Scraping Liability," Perkins Coie LLP, August 7, 2020, https://www.perkinscoie.com/images/content/2/3/236192/Law360-CFAA-Decision-May-Raise-Bar-On-Scraping-Liability.pdf.
[43] *FTC v. Roca Labs, Inc.,* 345 F. Supp. 3d 1375 (M.D. Fla. 2018); *McAfee v. State of New York*, 149 N.Y.S.2d 547 (N.Y. Misc. 1956).
[44] 15 USC §45b.
[45] Mark Zuckerberg, Update on Cambridge Analytica, Facebook, March 21, 2018, https://www.facebook.com/zuck/posts/10104712037900071; Alvin Chang, "The Facebook and Cambridge Analytica scandal, explained with a simple diagram," Vox, May 2, 2018, https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

accessed the database for unauthorized purposes *did not* violate the CFAA.[46] Nevertheless, there remains the possibility that another judge looking at a different set of facts could determine that accessing a service in violation of its policies violates the CFAA, or state statutes that vary significantly in wording and scope.

3. Scraping Audit

Description:

In a scraping audit, a computer program extracts data, typically publicly available data, by repeatedly querying the algorithm and obtaining or otherwise observing the results. For example, Googlebot, Google's crawler that automatically discovers and scans websites to index in its search engine, is one of the most prolific web crawlers on the internet. Scraping is generally done by using automated scraping tools, such as a browser extension, that can accomplish specifically what the user asks it to do (such as collecting all the images in a publicly accessible website).

There are certain standards that are put in place to facilitate interactions between websites and bots. The "robot exclusion standard"— also known as "robots.txt"—allows the operator of a website to indicate whether, and to what extent, the bots can scan the website.[47] However, the robots.txt signal is only a signal; whether this request not to be scanned has any legal effect depends on jurisdiction. In most jurisdictions, the law is unclear.[48] In practice, there are plenty of bots on the internet that disregard the robots.txt standard completely.[49]

There can also be some overlap between scraping and the crowdsourced audit, which can sometimes differentiate based on whether or not there was user consent to data collection. Researchers at New York University created a browser plug-in called "Ad Observer" that attempts to study advertisements featuring political content and misinformation on Facebook and YouTube.[50] The platform users could opt-in to the study by adding the plug-in to their browser, which allowed the research group to scrape advertisements seen on the users' newsfeed. The results were then aggregated in an effort to learn how ads are targeted on the Facebook platform.

Practical Limitations:

Platforms may try to prevent researchers from scraping their sites. NYU's Ad Observer collected information only about the advertisement, including the information Facebook gives about why the ad was targeted to that particular user, who the advertiser is, and the advertisement itself. It

---

[46] *Van Buren v. United States*, 206 L. Ed. 2d 822 (D.D.C. 2020). Previously, at least one lower court (*Sandvig v. Barr*, U.S. District Court for the District of Columbia) concluded that the CFAA does not criminalize violations of ToS, because criminalizing constitutionally protected speech that happens to violate a ToS would be a serious threat to the First Amendment. Naomi Gilens and Jamie Williams, "Federal Judge Rules It Is Not a Crime to Violate a Website's Terms of Service," Electronic Frontier Foundation, April 6, 2020, https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service.

[47] Essentially, a website owner can place a text file in the root of the website hierarchy in a particular format that signals to the bot where it is allowed to scan, if allowed at all. "About robots.txt," https://www.robotstxt.org/robotstxt.html.

[48] "Can a /robots.txt be used in a court of law?" https://www.robotstxt.org/faq/legal.html.

[49] Rachel Costello, "Robots.txt," Deepcrawl, https://www.deepcrawl.com/knowledge/technical-seo-library/robots-txt/.

[50] Ad Observer, NYU Cybersecurity for Democracy, https://adobserver.org.

did not share any identifiable information about the user or their friends.[51] Nevertheless, in August 2021, Facebook disabled the accounts of the researchers conducting the study, effectively halting their research.[52] Critics of the move suggested that Facebook was concerned the researchers could use the tool to gain insight into how Facebook's ad-targeting algorithm works, how the company utilizes users' information to target advertisements, and how its algorithms contribute to misinformation.[53]

Legal Limitations:

As with crowdsourced audits, non-technical access restrictions such as contracts (like a terms of service) could also be used to chill algorithmic audits. In 2022, the Ninth Circuit ruled in *HiQ Labs, Inc. v. LinkedIn Corp*. that accessing information on publicly available websites—or accessing information behind a technological barrier when the user is given authorization—is not a violation of the CFAA.[54] While this decision is good news for AI reseachers and auditors seeking to identify discriminatory outcomes or other harmful effects of algorithms, it may not extend to scraping of other non-public data sets to which a user has legitimate access.

An auditor scraping a public website without permission or in contravention to a robots.txt signal opting out of scraping could potentially be liable for common law trespass to chattels (or property) as well. Initially, some courts held that trespass to chattels can be a viable way to claim injury due to scraping, if there is demonstrable harm to the host computer or network. Generally, this term means an owner can claim injury if someone uses their property without the owner's permission; in the case of computers, the "property" can refer to a computer system or network.[55] In *eBay, Inc. v. Bidder's Edge, Inc*., eBay successfully argued that, while Bidder's Edge's spidering activity minimally harmed eBay's systems, a preliminary injunction could discourage more companies from doing the same—to not do so would encourage other companies to use web crawlers, hurting eBay's servers with this increased use of activity.[56] Other courts have since been more skeptical. In *Ticketmaster Corp. v. Tickets.com, Inc.,* a court held that scraping information from a public website on its own was not sufficient to show the physical injury to the host computer or network required in a trespass action, stating: "This court respectfully disagrees with other district courts' finding that mere use of a spider to enter a publically available website to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels."[57] There are, of course, legitimate reasons why a website owner would choose to not allow bots or other crawlers to access its web pages. Excessive bots can create high website traffic, which can strain servers and hurt the website's performance. Certain websites could also set prices for their products depending on traffic, so this could be

---

[51] Id.
[52] The research project continues to provide a searchable database of ads but has not disclosed from where it is receiving data. Mark Scott, "Fight over online political ads heats up ahead of midterms," Politico, August 3, 2022, https://www.politico.com/news/2022/08/03/2022-midterms-online-political-ads-00049373.
[53] Barbara Ortutay, "Facebook shuts out NYU academics' research on political ads," AP News, August 4, 2021, https://apnews.com/article/technology-business-5d3021ed9f193bf249c3af158b128d18.
[54] *HiQ Labs, Inc. v. LinkedIn Corp.*, 938 F.3d 985 (9th Cir. 2019).
[55] "Trespass to Chattels," Internet Law Treatise: Electronic Frontier Foundation, https://ilt.eff.org/Trespass_to_Chattels.html.
[56] *eBay, Inc. v. Bidder's Edge, Inc*., casebriefs.com, *https://www.casebriefs.com/blog/law/intellectual-property-law/intellectual-property-keyed-to-merges/state-intellectual-property-law-and-federal-preemption/ebay-inc-v-bidders-edge-inc*.
[57] *Ticketmaster Corp. v. Tickets.com, Inc.*, 2003 WL 21406289 (C.D. Cal. Mar. 7, 2003).

harmful to consumers if bot traffic artificially drives up prices (although this may be something a consumer-focused researcher would want to examine). Furthermore, scraping (using a bot or otherwise) can sometimes be harmful if a company chooses to use the data for potentially offensive purposes. Clearview AI, a controversial company that sells facial recognition tools to law enforcement, obtained billions of images used to train its models to identify individuals by scraping social media platforms and is now facing legal action from multiple governments.[58]

Because the design of a website or the content it contains may be copyrightable, when a researcher scrapes (copies) a website's content or information (for example, artwork for testing or reverse engineering an image processing algorithm), those researchers may open themselves up to liability for copyright infringement litigation. In *Ticketmaster Corp. v. Tickets.com, Inc.*, for instance, the court accepted that Ticketmaster's website was copyrightable but determined that Tickets.com spidering activity was fair use.[59] Fair use is a doctrine of U.S. copyright law allowing that the use of a copyrighted work "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright."[60] However, fair use is not a foolproof fail-safe, because the potential for high litigation costs to determine whether or not the use of a copyrighted work constituted a fair use can be a significant barrier for under-resourced or risk-averse entities likely to be conducting public interest research.

Even when using copyrighted material is considered fair use, the Digital Millennium Copyright Act prohibits the circumvention of technological measures that control access to copyright-protected works. This could include encryption systems, password-protected sections of websites, or digital rights management (DRM) software that is put in place to block access to copyrighted works—which includes software in which there is a copyright interest.[61] The law also prohibits the trafficking of tools put in place to help people circumvent these protection measures,[62] which could place in legal jeopardy researchers putting out APIs or other tools that allow individuals to audit algorithms.[63]

4. Sock Puppet Audit

Description:

In a sock puppet audit, a researcher creates fake accounts or programmatically constructed traffic for testing an algorithm. This gives the auditor control over each account's characteristics, making it easier to identify causality for discrimination or other harms. Another benefit is that

[58] "Clearview AI's unlawful practices represented mass surveillance of Canadians, commissioners say," Office of the Privacy Commissioner of Canada, February 3, 2021, https://www.priv.gc.ca/en/opc-news/news-and-announcements/2021/nr-c_210203/?=february-2-2021.
[59] *Ticketmaster Corp. v. Tickets.com, Inc.*, 2003 WL 21406289 (C.D. Cal. Mar. 7, 2003).
[60] 17 USC §107.
[61] Pub. L. 105-304.
[62] "Circumventing Copyright Controls," Digital Media Law Project: Berkman Klein Center for Internet and Society, September 10, 2021, https://www.dmlp.org/legal-guide/circumventing-copyright-controls.
[63] Consumer Reports has previously supported exemptions to section 1201 of the DMCA for good faith security research. https://advocacy.consumerreports.org/wp-content/uploads/2021/03/DMCA-13-expanding-security-research-3-9-21-FINAL-1.pdf.

auditors can assign characteristics to the fake accounts that volunteer participants might be hesitant to declare (such as medical history or sexual orientation).[64]

Practical Limitations:
Depending upon the nature of the study, the number of sock puppet accounts created may need to be quite large. This can be time-consuming and expensive, which is why semi-automated crowdsourcing like Amazon's Mechanical Turk is sometimes used for these studies.

Another drawback to this type of audit is that injecting large amounts of fake accounts into a system could tamper with the system in a way that interferes with the audit. For example, artificial traffic could drive up prices if a company notices there is high demand for a particular product.

Platforms that are designed to detect or deactivate fake accounts (or even identify third-party tests being done on their own system) could be able to remove these accounts before an audit is complete. This could be done to deliberately frustrate the audit or could simply be a result of standard efforts to detect and remove inauthentic accounts. Alternatively, a company could deliberately present different results to sock puppet accounts in order to present a better (and misleading) picture about the results generated by its algorithms.

Legal Limitations:
Similar to the previous auditing examples, breach of contract (if a ToS prohibits the creation of fake accounts, even for research purposes)[65] and trespass to chattels could be asserted against researchers creating fake accounts to conduct a sock puppet audit. Because platforms have legitimate reasons to monitor and delete fake accounts to avoid artificially inflated user counts or content promotion and to limit abuse of network resources, a court may be sympathetic to a legal challenge against even fake accounts created for auditing purposes.

In *Sandvig v. Barr*, academic researchers sought to study whether certain employment websites discriminated based on certain characteristics, and hoped to make fake accounts with these characteristics to examine how the platforms' algorithms behaved; however, this method violated many websites' terms of service. The researchers brought a pre-enforcement First Amendment challenge, alleging that the CFAA as applied to ToS violations chilled their free speech.[66] The court concluded that the CFAA does not criminalize violations of ToS, because criminalizing constitutionally protected speech that happens to violate a ToS would be a serious threat to the First Amendment.[67]

Policy Recommendations

Clearly, the legal and practical impediments to good faith public interest auditing are vast and could hinder research into identifying algorithmic harms. The various laws mentioned could

[64] Christian Sandvig et al., "Auditing Algorithms," 14.
[65] James Snell et al., "CFAA Decision May Raise Bar On Scraping Liability," Perkins Coie LLP.
[66] *Sandvig v. Barr*, 451 F. Supp. 3d 73 (D.D.C. 2020).
[67] Id.

pose a legal threat to auditors, preventing them from tinkering with algorithms for fear of legal recourse. We propose recommendations on ways to carve out exemptions to existing law to promote this research. Furthermore, we also provide recommendations on mandating data and code access in some cases to researchers to make model evaluation easier.

1. Access and Publication Mandates

Though code audits may not be necessary for lower-stakes applications of AI, for particularly sensitive applications, the code governing these decisions should be made available to the public, along with the training and testing data used. First, government uses of algorithms such as bail decisions in law enforcement and basic resource allocation should be transparent to the public, because these decisions impact people's liberties and basic rights (if these algorithms are to be used at all; a particular state bill would ban such sensitive algorithmic decision-making[68]). Disclosure of code or an API that researchers can use to test an algorithmic system should also be provided when it has the potential to affect the public in dangerous ways (for example, if an algorithm is pointing users to wrong or harmful information regarding public health).

Second, government agencies and their technology vendors should frequently update their publicly available code and datasets—whenever significant changes are made. Engineers are constantly testing and updating their algorithms, and datasets can often become outdated or updated to more accurately train models. For algorithms with significant legal effects, disclosure of code and training data would need to be published regularly to reflect changes.

As mentioned, the datasets used to train algorithms are also often necessary to properly audit those same algorithms—giving researchers access to just code may not be enough. Due to potential copyright infringement issues, we recommend a safe harbor for researchers using copies of AI training data for public interest purposes or that such use be considered fair use.

Platforms should put in place a process for researchers either to create fake accounts for auditing purposes or to appeal takedowns of research-related fake accounts. The platforms should also treat these accounts the same way they do their regular users; platforms should not be able to frustrate testing.[69] This may be difficult for smaller companies to implement but should be required for larger ones (determined by user count or annual revenue).

Finally, whether an algorithm is open-source or not could also be a factor to consider in assessing an AI designers' liability for discrimination, because transparency could be deemed a good faith effort at rooting out bad outcomes. Some companies choose to make their software open-source (or available to the public so that anyone can inspect, download, and test their code). While in some cases there could be a competitive disadvantage for a company to make its code public, there are numerous advantages in terms of reducing algorithmic bias and other harms. Anyone, including auditors, can inspect the code and test for issues—they can also

[68] S.B. 5116, 67th Legislature, 2021 Regular Session (Washington 2021), https://lawfilesext.leg.wa.gov/biennium/2021-22/Pdf/Bills/Senate%20Bills/5116.pdf?q=20210810140732.
[69] Frustrating testing for algorithmic harm/bias could be considered an unfair/deceptive practice or an unfair method of competition.

notify the company if anything concerning is found so that the company can fix it. Auditors can also provide code or other suggestions on how to improve the software.

2. CFAA and Computer Trespass

Recent decisions such as *HiQ* and *Van Buren* have found that users who had legitimate access to a computer service did not violate the CFAA when they exceeded the policy limitations imposed on such access.[70] This reduces the likelihood that the CFAA could be used against public interest researchers querying a database to test for bias, potentially in violation of a company's terms of service. In fact, the Department of Justice recently released a statement to federal prosecutors saying that it would not use the CFAA to prosecute good faith researchers attempting to identify security vulnerabilities.[71] While the DOJ did not mention whether this new policy would also apply to researchers of algorithmic bias and other harm, it could indicate the DOJ would be more hesitant to prosecute researchers working for the public good.

Nevertheless, the holdings of recent cases are necessarily limited to the fact patterns in question in those cases, and a court looking at a slightly different scenario could decide that the CFAA limits unwanted testing of an algorithm. Moreover, many of these decisions apply only to the Computer Fraud and Abuse Act itself: There may exist potential causes of action under comparable state statutory law or common law trespass to chattels. Policymakers should consider targeted reforms of these laws to ensure that good faith public interest research that does not meaningfully tax a company's resources or compromise other interests (such as privacy) is allowed—even for public-facing sites that use a robots.txt flag.

3. Contract Law

Today, many companies put language into terms of service or license agreements purporting to limit researchers' ability to access their systems to test for bias or other problems. Even if such clauses do not trigger the Computer Fraud and Abuse Act, they could still be the basis for private litigation against a user. At the very least, the threat of such a lawsuit could serve to deter audits that could uncover serious problems.

Under existing contract law, courts may determine that such clauses are unconscionable and void as against public policy.[72] However, that possibility does not provide certainty to risk-averse researchers who are likely to lack the resources to litigate against a large tech company.

Legislators should consider enacting legislation that explicitly prohibits contractual language unfairly limiting researchers' ability to audit algorithms for bias. Policymakers regularly pass laws

[70] Andrew Crocker, "Scraping Public Websites (Still) Isn't a Crime, Court of Appeals Declares," Electronic Frontier Foundation, April 19, 2022, https://www.eff.org/deeplinks/2022/04/scraping-public-websites-still-isnt-crime-court-appeals-declares.
[71] https://www.justice.gov/opa/press-release/file/1507126/download.
[72] "Contracts Considered to be Contrary to Public Policy," UpCounsel, https://www.upcounsel.com/what-contracts-are-considered-to-be-contrary-to-public-policy; Paul Bennett Marrow, "Contractual Unconscionability: Identifying and Understanding Its Potential Elements," Columbia.edu, 2000.

prohibiting the use of clauses that violate public policy interests: California, for example, prohibits noncompete clauses in employment contracts,[73] and President Biden recently signed a law that prohibits mandatory arbitration for sexual harassment claims, as well as claims of retaliation resulting from internal complaints of sexual assault or harassment.[74]

In 2016, Congress passed the Consumer Review Fairness Act, which bans contractual clauses that limit a consumer's ability to post honest reviews about a company online. However, this law does not explicitly cover clauses that limit the underlying testing that could lead to a negative review. To better facilitate transparency and accountability, the protections in this law could be extended to ban anti-testing clauses as well.

4. DMCA

The Library of Congress may create temporary exemptions every three years from the anti-circumvention provisions of Section 1201(a) for specified purposes, such as reverse engineering for security research.[75] The security research exemption might be read to encompass scraping to access works for algorithmic bias or harm testing for particular applications with significant legal effects. If not, an exemption to that effect should be proposed to the Copyright Office in the next Triennial Review, due to begin in mid-2023, and to conclude with new exemptions in late 2024. Or Congress could codify a new exemption in the statute itself.

5. Copyright

If the database underlying the development of an algorithm is copyrightable, then the unlicensed use of those works for algorithmic auditing should be considered fair use. Ultimately, researchers should not have to worry about whether the data they scrape in order to reverse engineer and train or test algorithms to identify harms leads to penalties from copyright infringement.[76]

6. Civil Rights, Privacy, and Security

Consumer Reports has long supported comprehensive privacy and security legislation to protect consumers.[77] Privacy and security rules should apply to public interest audits as well. While

[73] "Attorney General Bonta Reminds Employers and Workers That Noncompete Agreements Are Not Enforceable Under California Law," Press Release From CA Attorney General, March 15, 2022, https://oag.ca.gov/news/press-releases/attorney-general-bonta-reminds-employers-and-workers-noncompete-agreements-are.
[74] Public Law no: 117-90. Text: https://www.congress.gov/117/plaws/publ90/PLAW-117publ90.pdf.
[75] "Section 1201 Exemptions to Prohibition Against Circumvention of Technological Measures Protecting Copyrighted Works," U.S. Copyright Office, https://www.copyright.gov/1201/2021.
[76] "More Information on Fair Use," U.S. Copyright Office, https://www.copyright.gov/fair-use/more-info.html.
[77] Maureen Mahoney and Justin Brookman, "Consumer Reports Model State Privacy Act," Consumer Reports Digital Lab, February 2021, https://advocacy.consumerreports.org/wp-content/uploads/2021/02/CR_Model-State-Privacy-Act_022321_vf.pdf.

there is clear societal value to such research, that does not mean that researchers should have unfettered access to private data stores. Research exceptions to privacy laws should be narrowly tailored to be consistent with reasonable consumer expectations, and new access mandates to facilitate public interest research should limit third-party access to identifiable information. To the extent possible, data should be deidentified and aggregated before being handed over, and researchers should generally be prohibited from secondary use or sharing of data obtained for auditing purposes.

New privacy law should also include civil rights provisions that update decades-old protections to account for technologies such as artificial intelligence. Today civil rights protections are governed by different sector-specific statutes, each with its own standards and interpretations that have evolved over the years. However, in many cases, it is not clear how these protections apply when discriminatory outcomes are driven by a machine learning algorithm instead of by a conscious choice on the part of a company. Privacy legislation should comprehensively provide that discrimination that results in a loss of economic opportunities or access to public accommodations for members of protected classes is prohibited.[78] Bills like the recently introduced American Data Privacy and Protection Act take into account civil rights and algorithms, but the U.S. has yet to pass federal data privacy legislation.[79]

7. Consumer Protection Law

General purpose consumer protection law prohibits companies from engaging in "deceptive practices." Most deception cases are predicated on a company deceiving *a consumer*—such as lying about product attributes or misstating fees. However, other types of deceptive behavior can harm the marketplace and result in consumers being misled.

Companies that become aware they are subject to a public interest audit may make the decision to feed testers inaccurate information in order to paint a positive but misleading picture. Volkswagen famously settled after installing defeat devices[80] in certain diesel vehicles to detect when a car was being operated in a test environment in order to change pollution levels.[81] A third-party testing service has accused a cell phone manufacturer of engaging in similar tactics to game benchmarking tests.[82] An algorithm developer being tested for bias could try to detect auditors testing for bias and send them cleansed results reflecting an inaccurate depiction of normal results.

Currently the law is not entirely clear as to when deceiving third-party testers is illegal. The Federal Trade Commission settled a multibillion dollar case with Volkswagen, but its deception claims were based on deceiving consumers as to the environmental impact of its diesel

[78] "Consumer Reports Model State Privacy Act," 12.
[79] American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022).
[80] The EPA defines a defeat device as "any device that bypasses, defeats, or renders inoperative a required element of the vehicle's emission control system," https://www.epa.gov/vw/learn-about-volkswagen-violations.
[81] "Volkswagen Clean Air Act Civil Settlement," Environmental Protection Agency, https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement.
[82] Chris Smith, "Geekbench bans Galaxy S22 for cheating in benchmark tests," BGR, March 7, 2022, https://bgr.com/tech/geekbench-bans-galaxy-s22-for-cheating-in-benchmark-tests.

engines, not that Volkswagen deceived testers.[83] The FTC's Policy Statement on Deception—an informal but influential explanation of how the FTC interprets its legal authority—says that to allege deception, "there must be a representation, omission or practice that is likely to mislead *the consumer"* (emphasis added).[84] The FTC should update this nearly 40-year-old guidance to account for other forms of deception, and otherwise clarify to companies that providing misleading test results is actionable under the law.

[83] "In Final Court Summary, FTC Reports Volkswagen Repaid More Than $9.5 Billion To Car Buyers Who Were Deceived by 'Clean Diesel' Ad Campaign," Federal Trade Commission Press Release, July 27, 2020, https://www.ftc.gov/news-events/news/press-releases/2020/07/final-court-summary-ftc-reports-volkswagen-repaid-more-95-billion-car-buyers-who-were-deceived-clean. The FTC also alleged that Volkswagen's behavior was "unfair" to consumers because they were induced to purchase vehicles with a lower-than-expected resale value. Regulators could potentially bring similar unfairness claims against other companies that deceive testers, resulting in consumers purchasing products with less-than-expected functionality. However, to prove unfairness, regulators typically must allege elements—such as "substantial injury"—and that those harms were not offset by countervailing benefits. Further, many consumer protection regulators do not have unfairness authority; they can only bring deception cases. As such, deception should be available to regulators as a tool to proceed against companies that evade third-party auditing.

[84] "FTC Policy Statement on Deception," Federal Trade Commission, October 14, 1983, https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf.

Other Frameworks to Incentivize Public Interest Audits

Bug Bounty Programs for Algorithms

Bug bounty programs have previously been used by many websites and other software companies to identify and fix security vulnerabilities.[85] Generally, these companies offer compensation and recognition to individuals who can identify these vulnerabilities.

Companies like Twitter have been using this process to let the public identify issues with certain algorithms the platform uses. Twitter recently received backlash when it was discovered that its image-cropping algorithm, which showed previews of images and videos people tweeted, was shown to be biased toward younger, slimmer, and lighter faces.[86] For its algorithmic bias bug bounty program, the company released its code for this specific image-cropping algorithm and asked that individuals identify and taxonomize the potential harms that an algorithm like this can produce.[87]

However, Twitter's bug bounty program addressed only one algorithm used on the platform—the image-cropping algorithm is not the root cause of some of the major algorithmic problems that the platform continues to host, such as opaque content moderation practices, amplification of misinformation on the platform, and harmful advertisement delivery to users. It is unlikely that Twitter would publicly release the code to these algorithms that are central to its business, but allowing researchers this access would obviously be a more transparent way for the public to understand how these problems arise and might force Twitter to address these issues.

These platforms should allow the public to view their code and tackle some of their larger problems in exchange for reduced liability for potential harms if they act in good faith. Bounty programs should be considered relevant when assessing whether a company has met its obligations to root out bias or other algorithmic harm. However, companies will always have the best and most sophisticated view into their own systems; companies cannot simply punt their own obligations to assess systems for bias to the public via bounty programs.

---

[85] In December 2021, the Department of Homeland Security (DHS) announced a bug bounty program to identify potential cybersecurity vulnerabilities in certain DHS systems. Cybersecurity researchers were vetted to gain access to certain external DHS systems in order to find vulnerabilities and be compensated for the bugs they identify; "DHS Announces 'Hack DHS' Bug Bounty Program to Identify Potential Cybersecurity Vulnerabilities," U.S. Department of Homeland Security, December 14, 2021, https://www.dhs.gov/news/2021/12/14/dhs-announces-hack-dhs-bug-bounty-program-identify-potential-cybersecurity.

[86] Alex Hern, "Student proves Twitter algorithm 'bias' toward lighter, slimmer, younger faces," The Guardian, August 10, 2021, https://www.theguardian.com/technology/2021/aug/10/twitters-image-cropping-algorithm-prefers-younger-slimmer-faces-with-lighter-skin-analysis.

[87] "Twitter Algorithmic Bias," HackerOne, https://hackerone.com/twitter-algorithmic-bias?type=team.

Whistleblower Protections

Whistleblowing has the potential to be an effective way for employees to enact changes on company practices, which can include mitigating harmful algorithms. Due to the general lack of requirements that are placed on companies to be transparent about algorithmic bias, whistleblowers can often expose problems to the public that companies have no real incentive to disclose or address—particularly when the disclosure of such information could harm profits. In 2020, Google effectively forced out a top AI ethics researcher for trying to publish a paper critiquing the kinds of algorithms (large language models) that Google uses. The paper pointed out some of the harms that can come from these models, as well as other ethical considerations concerning these algorithms.[88] The conclusions of the paper itself were not entirely novel. However, this resulting controversy has led to suspicions that the creation of ethics teams within private companies may be little more than a PR stunt and that these teams do not necessarily have sway in terms of internal engineering practices and the products themselves.

It is clear that many AI companies cannot be trusted to always regulate themselves or be forthcoming about the issues in their algorithms. Whistleblowers can play an important role in providing the public and regulators with some clarity about how algorithms work and their associated impacts, particularly when companies perhaps know what the issues are but choose not to disclose or address these problems. Today, there are few protections given to whistleblowers in terms of disclosing issues related to AI. We will outline some potential policy changes that can provide some protections to whistleblowers while being fair to companies that are attempting to address discriminatory impacts of their products in good faith.

We recommend enacting protections for whistleblowers who attempt to disclose anything from algorithmic bias against protected classes to flawed research methodologies or data collection practices to false claims made by the company about its products. Individuals who bring up these issues internally to upper management if the company does not adequately address them within a certain time period, or for deployed models where potential discrimination is already in effect, should be protected from retaliation. This would include prohibiting whistleblowing in particular cases from affecting the employee's job status and prospects for promotion.

We also favor an approach that affirmatively incentivizes and protects whistleblowing (generally in the form of awards). As models, the Whistleblower Protection Enhancement Act (WPEA) protects federal employees who report fraud and abuse,[89] and the False Claims Act's qui tam provision protects anyone with evidence of fraud against federal programs or contracts and has awards for doing so. The IRS also has a whistleblower award for those who report on individuals who fail to pay the taxes they owe.[90] Other examples include the Sarbanes-Oxley Act, which provides whistleblower protections at public companies to encourage fraud reporting,

[88] Khari Johnson, "AI ethics pioneer's exit from Google involved research into risks and inequality in large language models," VentureBeat, December 3, 2020, https://venturebeat.com/2020/12/03/ai-ethics-pioneers-exit-from-google-involved-research-into-risks-and-inequality-in-large-language-models.
[89] "Whistleblower Information," U.S. CPSC Office of Inspector General, https://oig.cpsc.gov/whistleblower-information.
[90] "Whistleblower Office," Internal Revenue Service, https://www.irs.gov/compliance/whistleblower-office.

and to some extent the protections apply to private companies if they provide services for publicly traded ones.[91] Senator Brian Schatz (D-Hawaii) and Senator John Thune (R-S.D.) introduced the Platform Accountability and Consumer Transparency (PACT) Act in 2020, which would require the Government Accountability Office to study and report on the viability of an FTC-administered whistleblower and awards program for employees or contractors of online platforms.[92]

We also recommend prohibiting companies from forcing employees to sign nondisclosure or non-disparagement agreements regarding algorithmic bias or other unfair practices or outcomes regarding their company's technology. As a reference, California's Senate Bill 331, "The Silenced No More Act," adopted in 2021, prohibits workers from being forced to sign NDAs regarding all forms of worker discrimination and harassment in the workplace[93] (previous law in CA addressed only sexual harassment).

Furthermore, copyright law could hinder whistleblowers from publicly posting data or other information about algorithms. If an employee wanted to post a dataset their company was using to indicate its issues, this could be copyright infringement if the data itself was protected by copyright (for example, if the dataset contained artwork). We recommend that whistleblowers making copyrighted data related to algorithms publicly available for the purposes of disclosing its harmful effects should be considered a fair use case.

Conclusion

Certain applications of AI have the potential to roll back much of the progress made by civil rights law. Due to the lack of transparency on how these algorithms are used, the data used to train them, and how engineers go about mitigating harm when designing these algorithms, many of these algorithms may very well be discriminating against protected classes and perpetuating other kinds of harm. While the burden must not fall entirely on public interest researchers to uncover algorithmic harm, we must clear the legal barriers that hinder important public interest research as we advocate for robust algorithmic regulation in the U.S.

[91] Sarbanes-Oxley Act, 18 U.S.C. §1514A.
[92] PACT Act, S. 4066, 116th Cong. (2020).
[93] S.B. 331, California State Senate, 2021 Reg. Sess., (Cal. 2021), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220SB331; "California Silenced No More Act," Silenced No More Foundation, https://silencednomore.org/the-silenced-no-more-act.