Combatting Online Harms Through Innovation

Federal Trade Commission | Report to Congress

FEDERAL TRADE COMMISSION

June 16, 2022

COMBATTING ONLINE HARMS THROUGH INNOVATION

Federal Trade Commission

Report to Congress

June 16, 2022

Table of Contents

I. INTRODUCTION

In the 2021 Appropriations Act, Congress directed the Federal Trade Commission to study and report on whether and how artificial intelligence (AI) "may be used to identify, remove, or take any other appropriate action necessary to address" a wide variety of specified "online harms."[1] Congress refers specifically to content that is deceptive, fraudulent, manipulated, or illegal, and to particular examples such as scams, deepfakes, fake reviews, opioid sales, child sexual exploitation, revenge pornography, harassment, hate crimes, and the glorification or incitement of violence. Also listed are misleading or exploitative interfaces, terrorist and violent extremist abuse of digital platforms, election-related disinformation, and counterfeit product sales. Congress seeks recommendations on "reasonable policies, practices, and procedures" for such AI uses and on legislation to "advance the adoption and use of AI for these purposes."[2]

AI is defined in many ways and often in broad terms.[3] The variations stem in part from whether one sees it as a discipline (e.g., a branch of computer science), a concept (e.g., computers performing tasks in ways that simulate human cognition), a set of infrastructures (e.g., the data and computational power needed to train AI systems), or the resulting applications and tools.[4] In a broader sense, it may depend on who is defining it for whom, and who has the power to do so.[5]

---

[1] This language is from a section of the Act known as the American COMPETE Act, which directs the Commission to conduct a "Study to Combat Online Harms Through Innovation." Consolidated Appropriations Act, 2021, Pub. L. No. 116-260, Title XV, § 1501(j), https://www.govinfo.gov/app/details/BILLS-116hr133enr/summary.

[2] *Id.*

[3] For example, Congress has defined AI as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments." National Defense Authorization Act for Fiscal Year 2021, Div. E, § 5002(3), https://www.govinfo.gov/app/details/BILLS-116hr6395ih. The Organisation for Economic Co-operation and Development (OECD) has used the same definition. *See* OECD, *Recommendation of the Council on Artificial Intelligence* (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. The National Security Commission on Artificial Intelligence (NSCAI) defined it as the "ability of a computer system to solve problems and to perform tasks that have traditionally required human intelligence to solve." NSCAI Final Report (2021) at 659, https://www.nscai.gov/2021-final-report/. *See also* EDRi, *Beyond Debiasing: Regulating AI and Its Inequalities* at 22 (2021) (defining AI as a "broad set of computational methods that serve to perform a wide range of tasks automatically"), https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf; Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 UC Davis L. Rev. 399, 404 (2017) ("AI is best understood as a set of techniques aimed at approximating some aspect of human or animal cognition using machines"), https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Calo.pdf.

[4] *See, e.g.*, Matt Chessen, *What is Artificial Intelligence? Definitions for policy-makers and non-technical enthusiasts* (2017), https://medium.com/artificial-intelligence-policy-laws-and-ethics/what-is-artificial-intelligence-definitions-for-policy-makers-and-laymen-826fd3e9da3b. Further, vendors and others often misuse the term in their marketing efforts. *See* Frederike Kaltheuner, *This book is an intervention*, in Fake AI (Frederike Kaltheuner, ed.) (2021), https://meatspacepress.com/.

[5] That power is mostly in the hands of big technology companies. *See* Emily Tucker, *Artifice and Intelligence*, Tech Policy Press (Mar. 17, 2022), https://techpolicy.press/artifice-and-intelligence/. A global group of experts recently stated: "'Artificial' and 'intelligence' are loaded terms, their definitions subject to cultural biases. AI is a technology, a science, a business, a knowledge system, a set of narratives, of relationships, an imaginary." *See* AI

We assume that Congress is less concerned with whether a given tool fits within a definition of AI than whether it uses computational technology to address a listed harm. In other words, what matters more is output and impact.[6] Thus, some tools mentioned herein are not necessarily AI-powered. Similarly, and when appropriate, we may use terms such as automated detection tool or automated decision system,[7] which may or may not involve actual or claimed use of AI. We may also refer to machine learning, natural language processing, and other terms that — while also subject to varying definitions — are usually considered branches, types, or applications of AI.

We note, too, that almost all of the harms listed by Congress are not themselves creations of AI and, with a few exceptions like deepfakes, existed well before the Internet. Greed, hate, sickness, violence, and manipulation are not technological creations, and technology will not rid society of them.[8] While social media and other online environments can help bring people together, they also provide people with new ways to hurt one another and to do so at warp speed and with incredible reach.[9]

No matter how these harms are generated, technology and AI do not play a neutral role in their proliferation and impact. Indeed, in the social media context, the central challenge of the Congressional question posed here should not be lost: the use of AI to address online harm is merely an attempt to mitigate problems that platform technology — itself reliant on AI — amplifies by design and for profit in accord with marketing incentives and commercial surveillance. Harvard University Professor Shoshana Zuboff has explained that platforms'

Decolonial Manyfesto, https://manyfesto.ai/index.html?s=03. Computer scientist and Mozilla Fellow Deborah Raji lamented that editors ask her to use the term "AI" so that people will supposedly understand her, when in fact it "means nothing, by design." Deborah Raji, Twitter Post (Sep. 23, 2021), https://twitter.com/rajiinio/status/1441018006390415361?s=03.

[6] *See* Kristian Lum and Rumman Chowdhury, *What Is an "Algorithm"? It Depends on Who You Ask*, MIT Tech. Rev. (Feb. 26, 2021) ("What matters is the potential for harm, regardless of whether we're discussing an algebraic formula or a deep neural network."), https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/.

[7] Rashida Richardson, a Northeastern University School of Law professor currently working as an Attorney Advisor at the FTC, has explored definitional issues and explained her preference for the term "automated decision systems." Rashida Richardson, *Defining and Demystifying Automated Decision Systems*, 81 Md. L. Rev. ___ (forthcoming 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3811708.

[8] *See* Erin Saltman, *Challenges in Combating Terrorism and Extremism Online*, Lawfare (Jul. 11, 2021) ("We can't simply algorithm our way out of the problem"), https://www.lawfareblog.com/challenges-combating-terrorism-and-extremism-online; Olivia Solon, *Inside Facebook's efforts to stop revenge porn before it spreads*, NBC News (Nov. 19, 2019) (quoting Radha Plumb's observation that bad actors will always "figure out how to hurt people in ways that are very hard to predict or prevent"), https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631. Both Dr. Saltman, Director of Programming at the Global Internet Forum to Counter Terrorism, and Ms. Plumb, Chief of Staff to the Deputy Secretary at the Department of Defense, used to work at Facebook. *See also* Aspen Institute, *Commission on Information Disorder Final Report* at 15, 18 (Nov. 15, 2021), https://www.aspeninstitute.org/publications/commission-on-information-disorder-final-report/.

[9] Samidh Chakrabati, the former head of Facebook's civic integrity team, has argued that, if the harm at issue is greater than it would be if the content were shared in a chronological feed, via email, or via a subscription-based method, then the platform must bear some responsibility. *See* Samidh Chakrabati, Twitter Post (Dec. 13, 2021), https://twitter.com/samidh/status/1470446900738285569.

engagement engines — powering human data extraction and deriving from surveillance economics — are the crux of the matter and that "content moderation and policing illegal content" are mere "downstream issues."[10] Platforms do use AI to run these engines, which can and do amplify harmful content. In a sense, then, one way for AI to address this harmful content is simply for platforms to stop using it to spread that content. Congress has asked us to focus here, however, not on the harm that big platforms are causing with AI's assistance but on whether anyone's use of AI can help address any of the specified online harms.

Out of scope for this report are the widely expressed concerns about the use of AI in other contexts, including offline applications. As Congress directed, we focus here only on the use of AI to detect or address the specified online harms. Nonetheless, it turns out that even such well-intended AI uses can have some of the same problems — like bias, discrimination, and censorship — often discussed in connection with other uses of AI.

The FTC's work has addressed AI repeatedly, and this work will likely deepen as AI's presence continues to rise in commerce. Two recent FTC cases — one against Everalbum and the other against Facebook[11] — have dealt with facial recognition technology.[12] Commissioner Rebecca Kelly Slaughter has written about AI harms,[13] as have FTC staff members.[14] A 2016 FTC report, *Big Data: A Tool for Inclusion or Exclusion?*, discussed algorithmic bias in depth.[15] The agency has also held several public events focused on AI issues, including workshops on dark patterns and voice cloning, sessions on AI and algorithmic bias at PrivacyCon 2020 and 2021, a hearing on competition and consumer protection issues with algorithms and AI, a FinTech Forum on AI and blockchain, and an early forum on facial recognition technology (resulting in a 2012 staff

[10] Shoshana Zuboff, *You Are the Object of a Secret Extraction Operation*, The New York Times (Nov. 12, 2021), https://www.nytimes.com/2021/11/12/opinion/facebook-privacy html?s=03.

[11] Although Facebook has changed its corporate name to Meta, we continue to use the name Facebook in this report because many cited sources, as well as events and issues discussed therein, were published before the name change.

[12] *See* https://www ftc.gov/news-events/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers and https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions. Important remedies in these cases and a related action include required deletions of certain models, algorithms, data, or other work product. *See id.*; https://www ftc.gov/news-events/press-releases/2019/12/ftc-grants-final-approval-settlement-former-cambridge-analytica.

[13] *See* Rebecca Kelly Slaughter, *Algorithms and Economic Justice*, Yale J. L. & Tech. (Aug. 2021), https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf.

[14] *See* https://www ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai and https://www ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms.

[15] *See* FTC, *Big Data: A Tool for Inclusion or Exclusion* (Jan. 2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

report).[16] Some of these matters and events are discussed in more detail in the 2021 FTC Report to Congress on Privacy and Security.[17]

Reflecting this subject's importance, in November 2021, Chair Khan announced that the agency had hired its first-ever advisors on artificial intelligence.[18] The FTC has also sought to add more technologists to its professional staff. The FTC is not primarily a science agency, however, and is not currently authorized or funded to engage in scientific research beyond its jurisdiction. The FTC has traditionally consisted of lawyers, investigators, economists, and other professionals specializing in enforcement, regulatory, educational,[19] and policy efforts relating to consumer protection and competition. Some other federal agencies and offices do engage in more sustained AI-related work, sometimes as a central part of their mission.

With these agency caveats in mind, it is important to recognize that only a few of the harms Congress specified fall within the FTC's mission to protect consumers from deceptive or unfair commercial conduct. Many others do not, such as criminal conduct, terrorism, and election-related disinformation. It is possible, however, that changes to platforms' advertising-dependent business models, including the incentives for commercial surveillance and data extraction, could have a substantial impact in those categories. Further, some disinformation campaigns are simply disguises for commercially motivated actors.[20] We did consult informally with relevant federal agencies and offices on some issues, including the Department of State, the Department of Homeland Security (DHS), the Defense Advanced Research Projects Agency (DARPA), and the National Artificial Intelligence Initiative Office.[21] Thus, although we discuss each harm Congress lists, we would defer to other parts of the government on the topics as to which they are much more engaged and knowledgeable.

---

[16] *See* https://www ftc.gov/news-events/events-calendar/bringing-dark-patterns-light-ftc-workshop; https://www.ftc.gov/news-events/events-calendar/you-dont-say-ftc-workshop-voice-cloning-technologies; https://www.ftc.gov/news-events/events-calendar/privacycon-2021; https://www.ftc.gov/news-events/events-calendar/privacycon-2020; https://www.ftc.gov/news-events/events-calendar/ftc-hearing-7-competition-consumer-protection-21st-century; https://www.ftc.gov/news-events/events-calendar/2017/03/fintech-forum-blockchain-artificial-intelligence; and https://www.ftc.gov/news-events/events-calendar/2011/12/face-facts-forum-facial-recognition-technology.

[17] *See* FTC Report to Congress on Privacy and Security (Sep. 13, 2021), https://www.ftc.gov/reports/ftc-report-congress-privacy-security.

[18] *See* https://www ftc.gov/news-events/press-releases/2021/11/ftc-chair-lina-m-khan-announces-new-appointments-agency (announcing the appointments of Professor Meredith Whittaker, Amba Kak, and Sarah Myers West).

[19] Helping people avoid online harm is central to the FTC's consumer education efforts. *See, e.g.*, https://www.consumer ftc.gov/topics/online-security.

[20] *See* Elise Thomas, *Conspiracy Clickbait: This One Weird Trick Will Undermine Democracy*, Institute for Strategic Dialogue (2022), https://www.isdglobal.org/isd-publications/conspiracy-clickbait-this-one-weird-trick-will-undermine-democracy/#. The FTC is familiar with this kind of trick, having previously sued a company that used a political survey as a front for illegal robocalls that pitched cruise line vacations. *See* https://www.ftc.gov/news-events/press-releases/2015/03/ftc-ten-state-attorneys-general-take-action-against-political.

[21] Online harms and AI issues also being a topic of great interest globally, we have taken note of some of the current efforts that international agencies and organizations are taking to address them.

The scope of the listed harms leads to a few other preliminary observations. First, while that scope is broad, Congress does not ask for a report covering all forms of online harm or the general problem of online misinformation and disinformation. Second, the wide variety of the listed harms means that no one-size-fits-all answers exist as to whether and how AI can or should be used to address them. In some cases, AI will likely never be appropriate or at least not be the best option. Many of the harms are distinct in ways that make AI more or less useful or that would make regulating or mandating its use more or less of a legal minefield. For example, both AI *and* humans have trouble discerning whether particular content falls within certain categories of harm, which can have shifting and subjective meanings. Moreover, while some harms refer to content that is plainly illegal, others involve speech protected by the First Amendment. To the extent a harm *can* be clearly defined, AI tools can help to reduce it, albeit with serious limitations and the caveat that AI will never be able to replace the human labor required to monitor and contend with these harms across the current platform ecosystem.

Finally, we note that Congress does not refer to who may be deploying these tools, requiring their use, or responsible for their outcomes. The use of AI to combat online harm is usually discussed in the context of content moderation efforts by large social media platforms. We do not limit the report strictly to that context, however, because the Congressional language does not mention "social media" at all and refers to "platforms" only in connection with terrorists and violent extremists. Governments and others may deploy these tools, too. Other parts of the online ecosystem or "tech stack" are also fair game, including search engines, gaming platforms, and messaging apps. That said, the body of the report reflects the fact that much of the research and policy discussion in this area focuses on social media, and for good reasons. These platforms and other large technology companies maintain the infrastructure in which these harms have been allowed to flourish, and despite mixed incentives to deal with those harms, they also control most of the resources to develop and deploy advanced mitigation tools.

II. EXECUTIVE SUMMARY

The deployment of AI tools intended to detect or otherwise address harmful online content is accelerating. Largely within the confines — or via funding from — the few big technology companies that have the necessary resources and infrastructure, AI tools are being conceived, developed, and used for purposes including combat against many of the harms listed by Congress. Given the amount of online content at issue, this result appears to be inevitable, as a strictly human alternative is impossible or extremely costly at scale.

Nonetheless, it is crucial to understand that these tools remain largely rudimentary, have substantial limitations, and may never be appropriate in some cases as an alternative to human judgment. Their use — both now and in the future — raises a host of persistent legal and policy concerns. The key conclusion of this report is thus that governments, platforms, and others must exercise great caution in either mandating the use of, or over-relying on, these tools even for the important purpose of reducing harms. Although outside of our scope, this conclusion implies that, if AI is not the answer and if the scale makes meaningful human oversight infeasible, we must look at other ways, regulatory or otherwise, to address the spread of these harms.

A central failing of these tools is that the datasets supporting them are often not robust or accurate enough to avoid false positives or false negatives. Part of the problem is that automated systems are trained on previously identified data and then have problems identifying new phenomena (e.g., misinformation about COVID-19). Mistaken outcomes may also result from problems with how a given algorithm is designed. Another issue is that the tools use *proxies* that stand in for some actual type of content, even though that content is often too complex, dynamic, and subjective to capture, no matter what amount and quality of data one has collected. In fact, the way that researchers classify content in the training data generally includes removing complexity and context — the very things that in some cases the tools need to distinguish between content that is or is not harmful. These challenges mean that developers and operators of these tools are necessarily reactive and that the tools — assuming they work — need constant adjustment even when they are built to make their own adjustments.

The limitations of these tools go well beyond merely inaccurate results. In some instances, increased accuracy could itself lead to other harms, such as enabling increasingly invasive forms of surveillance. Even with good intentions, their use can also lead to exacerbating harms via bias, discrimination, and censorship. Again, these results may reflect problems with the training data (possibly chosen or classified based on flawed judgments or mislabeled by insufficiently trained workers), the algorithmic design, or preconceptions that data scientists introduce inadvertently. They can also result from the fact that some content is subject to different and shifting meanings, especially across different cultures and languages. These bad outcomes may also depend on who is using the tools and their incentives for doing so, and on whether the tool is being used for a purpose other than the specific one for which it was built.

Further, as these AI tools are developed and deployed, those with harmful agendas — whether adversarial nations, violent extremists, criminals, or other bad actors — seek actively to evade and manipulate them, often using their own sophisticated tools. This state of affairs, often referred to as an arms race or cat-and-mouse game, is a common aspect of many kinds of new technology, such as in the area of cybersecurity. This unfortunate feature will not be going away, and the main struggle here is to ensure that adversaries are not in the lead. This task includes considering possible evasions and manipulations at the tool development stage and being vigilant about them after deployment. However, this brittleness in the tools — the fact that they can fail with even small modifications to inputs — may be an inherent flaw.

While AI continues to advance in this area, including with existing government support, all of these significant concerns suggest that Congress, regulators, platforms, scientists, and others should exercise great care and focus attention on several related considerations.

First, **human intervention** is still needed, and perhaps always will be, in connection with monitoring the use and decisions of AI tools intended to address harmful content. Although the enormous amount of online content makes this need difficult to fulfill at scale, most large platforms acknowledge that automated tools aren't good enough to work alone. That said, even extensive human oversight would not solve for underlying algorithmic design flaws. In any event, the people tasked with monitoring the decisions these tools make — the "humans in the loop" — deserve adequate training, resources, and protection to do these difficult jobs. Employers should also provide them with enough agency and time to perform the work and

should not use them as scapegoats for the tools' poor decisions and outcomes. Of course, even the best-intentioned and well-trained moderators will bring their own biases to the work, including a tendency to defer to machines ("automation bias"), reflecting that moderation decisions are never truly neutral. Nonetheless, machines should not be allowed to discriminate where humans cannot.[22]

Second, AI use in this area needs to be meaningfully **transparent**, which includes the need for it to be explainable and contestable, especially when people's rights are involved or when personal data is being collected or used. Some platforms may provide more information about their use of automated tools than they did previously, but it is still mostly hidden or protected as trade secrets. Transparency can mean many things, and exactly what should be shared with which audiences and in what way are all questions under debate. The public should have more information about how AI-based tools are being used to filter content, for example, but the average citizen has no use for pages of code. Platforms should be more open to sharing information about these tools with researchers, though they should do so in a manner that protects the privacy of the subjects of that shared data. Such researchers should also have adequate legal protection to do their important work. Public-private partnerships are also worth exploring, with due consideration of both privacy and civil liberty concerns.

Third, and intertwined with transparency, platforms and other companies that rely on AI tools to clean up the harmful content their services have amplified must be **accountable** both for their data practices and for their results. After all, transparency means little, ultimately, unless we can do something about what we learn from it. In this context, accountability would include meaningful appeal and redress mechanisms for consumers and others — woefully lacking now and perhaps hard to imagine at scale — and the use of independent audits and algorithmic impact assessments (AIAs). Frameworks for such audits and AIAs have been proposed, but many questions about their focus, content, and norms remain. Like researchers, auditors also need protection to do this work, whether they are employed internally or externally, and they themselves need to be held accountable. Possible regulation to implement both transparency and accountability requirements is discussed below, and the import of focusing on these goals cannot be overstated, though they do not stand in for more substantive reforms.

Fourth, **data scientists and their employers** who build AI tools — as well as the firms procuring and deploying them — are responsible for both inputs and outputs. They should all strive to hire and retain diverse teams, which may help reduce inadvertent bias or discrimination, and to avoid using training data and classifications that reflect existing societal and historical inequities. Appropriate documentation of the datasets, models, and work undertaken to create these tools is important in this regard. They should all be concerned, too, with potential impact and actual outcomes, even though those designing the tools will not always know how they will ultimately be used. Further, they should always keep privacy and security in mind, such as in their treatment of the training data. It may be that these responsibilities need to be imposed on

[22] As Dr. Chris Gilliard has framed it, "Automating that racist thing is not going to make it less racist." *See* Surveillance Killjoy, Twitter Post (Apr. 20, 2021), https://twitter.com/hypervisible/status/1384502881538166784.

executives overseeing development and deployment of these tools, not merely pushed as ethical precepts.

Fifth, platforms and others should **use the range of interventions** at their disposal, such as tools that slow the viral spread or otherwise limit the impact of certain harmful content. Automated tools can do more with harmful content than simply identify and delete it. These tools can change how platform users engage with content and are thus internal checks on a platform's own recommendation engines that result in such engagement in the first place.[23] They include, among other things, limiting ad targeting options, downranking, labeling, or inserting interstitial pages with respect to problematic content. How effective any of these tools are — and under what circumstances — is unknown and often still dependent on detection of particular content, which, as noted, AI usually does not do well. In any event, the efficacy of these tools needs more study, which is severely hindered by platform secrecy. AI tools can also help map and uncover networks of people and entities spreading the harmful content at issue. As a corollary, they can be used to amplify content deemed authoritative or trustworthy. Assuming confidence in who is making those determinations, such content could be directed at populations that were targets of malign influence campaigns (debunking) or that may be such targets in the future (prebunking). Such work would go hand-in-hand with other public education efforts.

Sixth, it is possible to give individuals the ability to use AI tools to limit their personal exposure to certain harmful or otherwise unwanted content. Filters that enable people, at their discretion, to block certain kinds of sensitive or harmful content are one example of such **user tools**. These filters may necessarily rely on AI to determine whether given content should pass through or get blocked. Another example is middleware, a tailored, third-party content moderation system that would ride atop and filter the content shown on a given platform. These systems mostly do not yet exist but are the topic of robust academic discussion, some of which questions whether a viable market could ever be created for them.

Seventh, to the extent that any AI tool intended to combat online harm works effectively and without unfair or biased results, it would help for smaller platforms and other organizations to have **access** to it, since they may not have the resources to create it on their own. As noted above, however, these tools have largely been developed and deployed by several large technology companies as proprietary items. On the other hand, greater access to such tools carries its own set of problems, including potential privacy concerns, such as when datasets are transmitted with the algorithm. Indeed, access to user data should be granted only when robust privacy safeguards are in place. Another problem is that the more widely a given tool is in use, the easier it will be to exploit.

Eighth, given the limitations on using AI to detect harmful content, it is important to focus on key complementary measures, particularly the use of **authentication tools** to identify the source

[23] Professor Olivier Sylvain, who is now the FTC's Senior Advisor on Technology, has noted that these "design tweaks" ultimately have limited efficacy because they "bump[] up against a far more compelling market incentive to hold and quantify consumer attention for advertisers." Olivier Sylvain, *Platform Realism, Informational Inequality, and Section 230 Reform*, Yale L.J. Forum 131 at 485 (Nov. 16, 2021), https://www.yalelawjournal.org/forum/platform-realism-informational-inequality-and-section-230-reform.

of particular content and whether it has been altered. These tools — which could involve blockchain, among other things — can be especially helpful in dealing with the provenance of audio and video materials. Like detection tools, however, authentication measures have limits and are not helpful for every online harm.

Finally, in the context of AI and online harms, any **laws or regulations require careful consideration.** Given the various limits of and concerns with AI, explicitly or effectively mandating its use to address harmful content — such as overly quick takedown requirements imposed on platforms — can be highly problematic. The suggestion or imposition of such mandates has been the subject of major controversy and litigation globally. Among other concerns, such mandates can lead to overblocking and put smaller platforms at a disadvantage. Further, in the United States, such mandates would likely run into First Amendment issues, at least to the extent that the requirements impact legally protected speech. Another hurdle for any regulation in this area is the need to develop accepted definitions and norms not just for what types of automated tools and systems are covered but for the harms such regulation is designed to address.

Putting aside laws or regulations that would require more fundamental changes to platform business models, the most valuable direction in this area — at least as an initial step — may be in the realm of transparency and accountability. Seeing and allowing for research behind platforms' opaque screens (in a manner that takes user privacy into account) may be crucial for determining the best courses for further public and private action.[24] It is hard to craft the right solutions when key aspects of the problems are obscured from view.

III. USING ARTIFICIAL INTELLIGENCE TO COMBAT ONLINE HARMS

A. Deceptive and fraudulent content intended to scam or otherwise harm individuals

Of the harms specified by Congress, deception is the most central to the Commission's consumer protection mission. Public and private sector use of AI tools to combat online scams is still in its relative infancy, and such tools may be hard to develop. While some scams may be detected by relatively clear and objective markers, many are context-dependent and not obvious on their face. After all, the nature of a scam is to deceive people into thinking it's not a scam. For example, the initial part of a scheme may involve a seemingly legitimate online ad, with key fraud indicators hidden offline and revealed only later. These factors may make it difficult for

[24] Commission staff is currently analyzing data collected from several large social media and video streaming companies about their collection and use of personal information as well as their advertising and user engagement practices. *See* https://www.ftc.gov/reports/6b-orders-file-special-reports-social-media-video-streaming-service-providers. In a 2020 public statement about this project, Commissioners Rebecca Kelly Slaughter and Christine S. Wilson remarked that "[i]t is alarming that we still know so little about companies that know so much about us" and that "[t]oo much about the industry remains dangerously opaque." https://www.ftc.gov/system/files/documents/public_statements/1584150/joint_statement_of_ftc_commissioners_chopra_slaughter_and_wilson_regarding_social_media_and_video.pdf.

machines to predict the veracity of claims about a product or service. [25] Automated tools may thus be less likely, at least in the near term, to help address online fraud as opposed to other harms.[26]

Despite the challenges, the use of AI to combat fraud is certainly an area for further research. Perhaps AI-based tools could help law enforcement agencies, researchers, platforms, and others reveal patterns of fraud and hidden connections between bad actors. A few consumer protection agencies have indeed started to look into whether AI can help in specific areas of fraud. For example, Japan's Consumer Affairs Agency sought funds to create an AI tool to identify websites selling products with deceptive COVID-19 prevention claims.[27] Poland's Office of Competition and Consumer Protection began a project to develop an AI tool that automatically detects unlawful clauses in business-to-consumer contracts.[28] Multiple Austrian government agencies have funded the development of an AI tool that would help consumers detect whether a website is a fake online shop.[29]

Facebook states that it uses AI tools to address various types of fraud, though, as the FTC has reported, scams on Facebook and other social media sites have continued to rise.[30] Specifically, Facebook says that it uses machine learning to identify scams and imposters on Messenger[31] and generally that it demotes content associated with fraud, including: links to suspected cloaking domains (which might involve financial scams); pages predicted to be spam (which might involve false ads, fraud, and security risks); and exaggerated health claims.[32] It also uses

---

[25] At the same time, scammers can use their own automated tools to commit fraud or can use the automated systems of social media platforms to amplify and target false advertising. *See*, *e.g.*, Jon Bateman, *Get Ready for Deepfakes to Be Used in Financial Scams*, Techdirt (Aug. 10, 2020), https://www.techdirt.com/2020/08/10/get-ready-deepfakes-to-be-used-financial-scams/; https://www ftc.gov/business-guidance/blog/2021/11/ftc-analysis-shows-covid-fraud-thriving-social-media-platforms; https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021.

[26] In contrast, AI is a common feature of anti-fraud measures in the credit card context, in which markers of fraud may be easier to detect. *See* PYMNTS and Brighterion, *AI in Focus: The Rise against Payments Fraud* (2021), https://www.pymnts.com/wp-content/uploads/2021/12/PYMNTS-AI-In-Focus-Waging-Digital-Warfare-Against-Payments-Fraud-December-2021.pdf; https://usa.visa.com/visa-everywhere/security/outsmarting-fraudsters-with-advanced-analytics html; https://www.paypal.com/us/brc/article/enterprise-solutions-paypal-machine-learning-stop-fraud. Payment firms also have the benefit of robust, accurate data about their customers and strong financial incentives to combat such fraud.

[27] *See* https://www.caa.go.jp/policies/budget/assets/policies_budget_201225_0002.pdf.

[28] *See* https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/org-details/949814883/project/899954/program/31061273/details.

[29] *See* https://www.ait.ac.at/en/news-events/single-view/detail/6860?cHash=250795314de77d44fa029af1a1310da2 and Louise Beltzung, et al., *Real-Time Detection of Fake-Shops through Machine Learning*, 2020 IEEE International Conference on Big Data (Dec. 2020), https://doi.org/10.1109/BigData50022.2020.9378204.

[30] *See* https://www ftc.gov/news-events/press-releases/2022/01/ftc-finds-huge-surge-consumer-reports-about-losing-money-scams; https://www ftc.gov/news-events/blogs/business-blog/2021/11/ftc-analysis-shows-covid-fraud-thriving-social-media; https://www ftc.gov/news-events/blogs/data-spotlight/2020/10/scams-starting-social-media-proliferate-early-2020.

[31] *See* https://messengernews fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/.

[32] *See* https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/.

automated detection systems for scams in the Facebook Marketplace, though their efficacy is uncertain at best.[33]

Other companies reporting similar usage of AI include Google, which uses it to detect online frauds and spam in search results and to filter spam, malware, and phishing in Gmail.[34] Microsoft makes similar use of AI for phishing and spam in Outlook.[35] Representatives of cybersecurity companies state that AI tools can also "track patterns in scam emails using large datasets and delete scam emails from people's inboxes."[36] Third-party vendors may offer AI-based scam detection as well, such as tools to find tax scam websites.[37]

Some academic research has also focused on the use of AI to address this harm, including a publicly funded project in the United Kingdom to detect fake dating profiles,[38] two connected studies on detecting undisclosed influencer affiliations,[39] and studies on detecting email spam.[40]

It is also worth noting that AI tools could aid in the investigation of whether companies are engaged in online conduct that harms competition. In 2021, for example, investigative journalists

---

[33] *See* Craig Silverman, et al., *Facebook Grew Marketplace to 1 Billion Users. Now Scammers Are Using It to Target People Around the World*, ProPublica (Sep. 22, 2021), https://www.propublica.org/article/facebook-grew-marketplace-to-1-billion-users-now-scammers-are-using-it-to-target-people-around-the-world. *See also* Colin Lecher and Surya Mattu, *Facebook Scammers Are Schilling Fake Cryptocurrency Using Big Tech's Biggest Names*, The Markup (Feb. 22, 2022), https://themarkup.org/citizen-browser/2022/02/22/facebook-scammers-are-schilling-fake-cryptocurrency-using-big-techs-biggest-names.

[34] *See* https://developers.google.com/search/blog/2021/04/how-we-fought-search-spam-2020 and https://security.googleblog.com/2020/02/improving-malicious-document-detection html?m=1.

[35] *See* https://www microsoft.com/en-us/insidetrack/office-365-helps-secure-microsoft-from-modern-phishing-campaigns; https://docs.microsoft.com/en-us/microsoft-365/security/office-365-security/anti-spam-protection?view=o365-worldwide. AlgorithmWatch has documented how some email spam filters using machine learning, including Microsoft's, may seem uncontroversial but could in fact be discriminatory and remain largely unaudited. Nicolas Kayser-Bril, *Spam filters are efficient and uncontroversial. Until you look at them*, AlgorithmWatch (Oct. 22, 2020), https://algorithmwatch.org/en/spam-filters-outlook-spamassassin/.

[36] *See* Social Catfish, *State of Internet Scams 2021* at 52-53, https://spcdnblog.socialcatfish.com/uploads/2021/07/State-of-Internet-Scams-2021-2.pdf.

[37] *See* Louise Matsakis, *Filing Your Taxes? Watch Out for Phishing Scams*, WIRED (Apr. 4, 2019), https://www.wired.com/story/filing-taxes-phishing-scams/.

[38] *See* https://webarchive.nationalarchives.gov.uk/ukgwa/20200930155453/https://epsrc.ukri.org/newsevents/news/aionlinedating/.

[39] *See* Arunesh Mathur et al., *Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest*, Proc. of the ACM on Human-Computer Interaction, Vol. 2, CSCW, Art. 119 (Nov. 2018), https://doi.org/10.1145/3274388; Michael Swart, et al., *Is This an Ad?: Automatically Disclosing Online Endorsements on YouTube with AdIntuition*, in CHI '20: Proc. of the 2020 CHI Conf. on Human Factors in Computing Systems (Apr. 2020), http://dx.doi.org/10.1145/3313831.3376178.

[40] *See* Emmanuel Gbenga Dada, et. al., *Machine Learning for Email Spam Filtering*, Heliyon 5(6) (2019), https://www.sciencedirect.com/science/article/pii/S2405844018353404#bib127.

for The Markup used a machine learning tool to examine whether "Amazon routinely ranked its own brands and exclusives ahead of better-known brands with higher star ratings."[41]

One caveat for consumer protection or competition enforcers, however, is that it makes little sense to use limited resources to obtain any AI tools without having already decided what exactly to do with them. It would be more sensible to determine first what an agency wants to find or learn and then see what available tools, AI or not, are best suited and most appropriate for that task. Of course, the agency would also need staff capable of deploying such tools and evaluating their responsible use.

B. Manipulated content intended to mislead individuals, including deepfake videos and fake individual reviews

Deepfakes

While most of the Congressionally specified harms predate and exist outside the online environment, deepfakes — and similar forms of synthetic media or media manipulation — are creatures of it. Deepfakes are video, photo, text, or audio recordings that seem real but have been manipulated with AI.[42] It would stand to reason, then, that AI or other sophisticated technology could help with — if not be integral to — detection of deepfakes. Indeed, public and private research on AI solutions to the deepfake problem have been underway for some time. As detection technology continues to improve, however, so will the ability to evade it, meaning that this AI battle will not end soon and that technological mitigation is insufficient.[43] A team overseen by DHS issued a report in 2021 that came to this conclusion and recommended pairing improved and constantly updated detection tools — to be used proactively and open-sourced as appropriate — with new laws, public-private cooperation, scientific responsibility, authentication tools, and education, with due consideration for civil liberties.[44]

---

[41] *See* Julia Angwin, *The Mathematics of Amazon's Advantage*, Hello World #79 (Oct. 16, 2021), https://www.getrevue.co/profile/themarkup/issues/the-mathematics-of-amazon-s-advantage-803575.

[42] In 2020, the Commission explored issues related to voice cloning, a subcategory of deepfakes, in a public workshop. *See* https://www.ftc.gov/news-events/events-calendar/you-dont-say-ftc-workshop-voice-cloning-technologies.

[43] *See* DHS, *Increasing Threat of Deepfake Identities* at 29 (Sep. 2021), https://www.dhs.gov/publication/2021-aep-deliverables; Government Accountability Office, *Science & Tech Spotlight: Deepfakes* (Feb. 2020), https://www.gao.gov/assets/gao-20-379sp.pdf; European Parliamentary Research Service ("EPRS"), *Tackling deepfakes in European policy* at 16-19, 60 (Jul. 2021), https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690039; Siwei Lyu, *Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms Race*, Scientific American (Jul. 20, 2020), https://www.scientificamerican.com/article/detecting-deepfakes1/; James Vincent, *Facebook develops new method to reverse-engineer deepfakes and track their source*, The Verge (Jun. 16, 2021), https://www.theverge.com/2021/6/16/22534690/facebook-deepfake-detection-reverse-engineer-ai-model-hyperparameters.

[44] DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 3, 29-34.

For several years, DARPA has been engaged intensively in studies on the automated detection of deepfakes and similarly manipulated content. These substantial efforts include the Semantic Forensics (SemaFor) program, led by Dr. Matt Turek, which seeks to develop tools "capable of automating the detection, attribution, and characterization of falsified media."[45] These last two factors are significant because, along with identifying that content has been manipulated, it is important both to know if it comes from where it claims to originate and to reveal the intent behind the manipulation.

Some big technology firms have been active in this area. The Deepfake Detection Challenge (DFDC), a joint effort involving the Partnership on AI (PAI), was a machine learning competition, launched initially on Facebook, created to incentivize development of technical means to detect AI-generated videos.[46] Separately, Google conducted an experiment, Assembler, that "aimed to advance how new detection technology could help fact-checkers and journalists identify manipulated media."[47] The DFDC and Assembler results reflect, among other things, the need for technology that can better recognize synthetic images in the wild (*i.e.*, in real-world circumstances) and not already in training datasets.[48] Facebook AI and Google AI both released their datasets for researcher use.[49] Many vendors offer tools for deepfake detection, too.[50]

Other research into detection methods, including the work of Professor Hany Farid at the University of California, Berkeley, is also well underway — in some cases with funding from the federal government and big technology companies — and reflects both promise and the need for continual attention and improvement.[51] Further, pursuant to its own studies of deepfake

---

[45] *See* https://www.darpa mil/news-events/2021-03-02.

[46] *See* Partnership on AI, *The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity* (Mar. 12, 2020), http://partnershiponai.org/wp-content/uploads/2021/07/671004_Format-Report-for-PDF_031120-1.pdf; https://ai facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

[47] *See* https://projectassembler.org.

[48] *See* Will Knight, *Deepfakes Aren't Very Good, Nor Are the Tools to Detect Them*, WIRED (Jun. 12, 2020), https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/; https://projectassembler.org/learnings/ (also noting unique challenges involving small or low-resolution images).

[49] *See* Brian Dolhansky, et al., *The DeepFake Detection Challenge (DFDC) Dataset* (Oct. 2020), https://arxiv.org/abs/2006.07397; https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection html. More recently, Facebook has open-sourced models developed pursuant to its Image Similarity Challenge, an attempt to advance at-scale detection of manipulated images. *See* https://ai.facebook.com/blog/detecting-manipulated-images-the-image-similarity-challenge-results-and-winners/?s=03.

[50] *See, e.g.*, https://weverify.eu/tools/deepfake-detector/; https://www fakenetai.com/; https://sensity.ai/deepfakes-detection/; https://www.mcafee.com/blogs/enterprise/security-operations/the-deepfakes-lab-detecting-defending-against-deepfakes-with-advanced-ai/; https://github.com/resemble-ai/resemblyzer; and https://cyabra.com/industries/.

[51] *See, e.g.*, Shruti Agarwal, et al., *Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches* (2020), https://farid.berkeley.edu/downloads/publications/cvpr20a.pdf, and *Detecting Deep-Fake Videos from Appearance and Behavior* (2020), https://farid.berkeley.edu/downloads/publications/wifs20.pdf; Luisa Verdoliva, *Media Forensics and DeepFakes: An Overview* (2020), https://arxiv.org/abs/2001.06564; Yuezun Li, et al., *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics* (2020), https://arxiv.org/abs/1909.12962; Andreas Rossler, et al., *FaceForensics++: Learning to Detect Manipulated Facial Images* (2019), https://arxiv.org/abs/1901.08971.

detection, PAI concluded that, to improve these tools, developers and deployers need to consider: the quality of detection models (which are not so good as to obviate the need for human review); how these models can be built outside of big platforms; how to agree on what manipulated media is harmful; and how to deal with low-tech manipulations ("cheapfakes") and misleading context.[52] A related challenge explored by PAI's Claire Leibowicz and others involves how to address the trade-offs resulting from adversarial dynamics. Specifically, effective detection tools should be available not only to big tech companies but also to smaller platforms, journalists, researchers, and others who can put them to good use — but the wider such tools are distributed, the easier it is for bad actors to defeat them.[53]

As noted above, while more research and development of detection methods should be encouraged, such technology will not be sufficient on its own. University of Texas Professor Robert Chesney, University of Virginia Professor Danielle Citron, and Professor Farid state that "[e]ven if capable detection technologies emerge … it is not assured that they will prove scaleable, diffusible and affordable to the extent needed to have a dramatic impact on the deepfake threat."[54] Similarly, a report from the Washington University's Center for an Informed Public concludes that a multi-stakeholder approach is necessary in part because "[t]he technology to detect deepfakes, and synthetic media more broadly, is imperfect, super hard to deliver at scale and speed, and still evolving."[55] Reflecting the state of the art in this area, a recent study showed that a leading deepfake detection model did no better than a group of ordinary people, though they made different kinds of mistakes.[56]

A separate, oft-cited concern raised by Professors Chesney and Citron involves what they coined the "Liar's Dividend," a dilemma arising from how increased public knowledge of deepfakes

---

[52] *See* Claire Leibowicz, et al., *Manipulated Media Detection Requires More Than Tools*, Partnership on AI (Jul. 13, 2020), https://www.partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/; Partnership on AI, *The Deepfake Detection Challenge*, *supra* note 46.

[53] *See* Claire Leibowicz, et al., *How to Share the Tools to Spot Deepfakes (Without Breaking Them)*, Partnership on AI Blog (Jan. 13, 2022), https://medium.com/partnership-on-ai/how-to-share-the-tools-to-spot-deepfakes-without-breaking-them-53d45cd615ac (discussing a framework for addressing the issue); Claire Leibowicz, et al., *The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media* (2021), https://arxiv.org/abs/2102.06109. *See also* Steven Prochaska, et al., *Deepfakes in the 2020 Elections and Beyond*, Wash. U Center for an Informed Public at 6-7 (Oct. 2020), https://cpb-us-e1.wpmucdn.com/sites.uw.edu/dist/6/4560/files/2020/10/CIP_Deepfake_Report_Summary-1.pdf; Sam Gregory, *The World Needs Deepfake Experts to Stem This Chaos*, WIRED (Jun. 24, 2021), https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/.

[54] Robert Chesney, Danielle Citron, and Hany Farid, *All's Clear for Deepfakes: Think Again*, Lawfare (May 11, 2020), https://www.lawfareblog.com/alls-clear-deepfakes-think-again.

[55] *See* Prochaska, *supra* note 53, at 1; James Vincent, *Deepfake detection algorithms will never be enough*, The Verge (Jun. 27, 2019), https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work; Henry Ajder and Nina Schick, *Deepfake apps are here and we can't let them run amok*, WIRED UK (Mar. 30, 2021) (noting that "no social media platform currently has deepfake detection in their media upload pipelines, and implementing detection on messaging apps like WhatsApp or Telegram would require monitoring users' conversations"), https://www.wired.co.uk/article/deepfakes-security.

[56] *See* Matthew Groh, et al., *Deepfake detection by human crowds, machines, and machine-informed crowds*, PNAS 119:1 (2022), https://doi.org/10.1073/pnas.2110013119.

makes it easier for people to escape accountability for their actions by denouncing authentic content as fake.[57] DHS has also identified this concern as a serious societal threat.[58] It is not a merely theoretical one,[59] and it is a conundrum somewhat analogous to how disseminating detection tools can lead inexorably to their speedier evasion.

Beyond the state of deepfake detection technology and its challenges exist questions about how those who possess that technology are using it. Different platforms may have different policies, with little transparency about their implementation and effect.[60] It is important to know, for example, how platforms determine the context of any given instance of manipulated media to ensure that artistic, satiric, and privacy-forward purposes are protected, and to be able to assess how well their systems work in making those distinctions. Such benign purposes are not theoretical, as manipulated media has a wide variety of legitimate uses.[61]

Given the many challenges of keeping detection technology at a level commensurate with deepfake technology, it is important to focus on the flip side: authentication. In other words, if it is difficult to identify fake content, then also try verifying real content.[62] Reflecting this pairing, Microsoft announced two new technologies in 2020 as part of its Defending Democracy Program: (1) Microsoft Video Authenticator, an AI-based deepfake detection tool; and (2) technology for its Azure cloud service and the BBC's Project Origin allowing content

---

[57] Robert Chesney and Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753, 1758 (2019) ("As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authentic video and audio as deep fakes"), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954.

[58] *See* DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 36.

[59] *See* Prochaska, *supra* note 53 at 9-10; Drew Harwell, *Top AI researchers race to detect 'deepfake' videos: 'We are outgunned*,' The Washington Post (Jun. 12, 2019), https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/.

[60] *See* Amber Frankland and Lindsay Gorman*, Combating the Latest Technological Threat to Democracy: A Comparison of Facebook and Twitter's Deepfake Policies,* German Marshall Fund (Jan. 13, 2020), https://securingdemocracy.gmfus.org/combating-the-latest-technological-threat-to-democracy-a-comparison-of-facebooks-and-twitters-deepfake-policies/; Harwell, *supra* note 59.

[61] Examples abound and include a weekly satire show, *Sassy Justice*, that used deepfakes as part of its premise. *See* Karen Hao, *The creators of South Park have a new weekly deepfake satire show*, MIT Tech. Rev. (Oct. 28, 2020), https://www.technologyreview.com/2020/10/28/1011336/ai-deepfake-satire-from-south-park-creators/. A documentary, *Welcome to Chechnya*, used deepfakes to protect LGBTQ people facing significant persecution. *See* Rebecca Heilweil, *How deepfakes could actually do some good*, Vox recode (Jun. 29, 2020), https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya.

[62] *See, e.g.*, DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 31; Prochaska et al., *supra* note 53, at 5-6. Alex Engler, Brookings Institution, *Fighting deepfakes when detection fails* (Nov. 11, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/; *National Security Challenges of Artificial Intelligence, Manipulated Media, and Deepfakes*, H. Perm. Select Comm. on Intelligence, 116th Cong. (2019) (testimony of Clint Watts), https://docs house.gov/meetings/IG/IG00/20190613/109620/HHRG-116-IG00-Wstate-WattsC-20190613.pdf; Verdoliva, *supra* note 51; Ashish Jaiman, *Technical Countermeasures to Deepfakes*, Towards Data Science (Aug. 27, 2020), https://towardsdatascience.com/technical-countermeasures-to-deepfakes-564429a642d3.

producers to add digital fingerprints to their content.[63] Like Adobe's Content Credentials, the latter would allow viewers, via browser extensions or other readers, to see the producer's identity and whether the content is authentic and unaltered.[64] Content authenticity goes beyond deepfakes, and broader efforts in this area, like those of the Coalition for Content Provenance and Authenticity, are discussed further below.

Fake reviews

The Commission has brought several lawsuits alleging fake or deceptive reviews of products and services, a subject that is essentially a subset of consumer fraud, discussed above.[65] This area remains a priority for the Commission, as evidenced by a recent Notice of Penalty Offenses Concerning Deceptive or Unfair Conduct around Endorsements and Testimonials that the FTC distributed to hundreds of businesses.[66]

Many platforms that feature reviews state that they use machine learning tools — usually in conjunction with some level of human review — to identify and remove fake reviews. The list includes large platforms like Google, Amazon, and Apple; review platforms like Yelp, TripAdvisor, and Trustpilot; and vendors like PowerReviews and BazaarVoice, which offer review-related services to major online retailers.[67] Further, several research papers, some developed with public funding, discuss the development of AI tools to detect fake reviews in different online environments, such as app stores.[68]

---

[63] *See* https://blogs microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/; Leo Kelion, *Deepfake detection tool unveiled by Microsoft*, BBC News (Sep. 1, 2020), https://www.bbc.com/news/technology-53984114.

[64] *See id.*; https://blog.adobe.com/en/publish/2021/10/26/adobe-unleashes-content-attribution-features-photoshop-beyond-max-2021#gs.kc3s0g.

[65] *See, e.g.*, https://www ftc.gov/news-events/press-releases/2022/01/fashion-nova-will-pay-42-million-part-settlement-ftc-allegations; https://www.ftc.gov/news-events/press-releases/2020/02/operators-comparison-shopping-website-agree-settle-ftc-charges; https://www.ftc.gov/news-events/press-releases/2019/02/ftc-brings-first-case-challenging-fake-paid-reviews-independent. Fake reviews are also the subject of FTC guidance for businesses and consumers. *See* https://www ftc.gov/reviews and https://www.consumer.ftc.gov/articles/how-evaluate-online-reviews.

[66] *See* https://www ftc.gov/enforcement/penalty-offenses/endorsements.

[67] *See* https://blog.google/products/maps/google-maps-101-how-we-tackle-fake-and-fraudulent-contributed-content/; https://www.aboutamazon.com/news/how-amazon-works/creating-a-trustworthy-reviews-experience; https://www.apple.com/newsroom/2021/05/app-store-stopped-over-1-5-billion-in-suspect-transactions-in-2020/; https://trust.yelp.com/recommendation-software/; https://www.tripadvisor.com/TripAdvisorInsights/w3690; https://cdn.trustpilot net/trustsite-consumersite/trustpilot-transparency-report-2021.pdf (at p. 26); https://www.powerreviews.com/blog/human-moderation-reviews/ (noting that all content is reviewed by human moderators after passing through automated filters); and https://knowledge.bazaarvoice.com/wp-content/conversations/en_US/Learn/moderation html.

[68] *See, e.g.*, Luis Gutierrez-Espinosa, et al., *Ensemble Learning for Detecting Fake Reviews*, 2020 IEEE 44th Annual Computers, Software, and Applications Conference (Jul. 2020), https://www.researchgate net/publication/345374735; Daniel Martens and Walid Maalej, *Towards understanding*

Some companies have developed their own, AI-based tools to detect suspicious reviews for the public. FakeSpot and ReviewMeta offer consumers insight into the authenticity of particular reviews on Amazon and other platforms, relying in part on machine learning tools.[69] In 2021, FakeSpot released a report stating that it had used AI to determine the extent of unreliable product reviews on the Amazon, Walmart, eBay, Best Buy, Shopify, and Sephora websites.[70] It found that, in 2020, nearly 31% of the reviews on those sites were unreliable, though the percentage varied significantly between sites, with Walmart faring the worst and Best Buy the best.[71] Similarly, a 2021 report by Uberall and The Transparency Company relied on machine learning to determine review authenticity on Google My Business (GMB), Facebook, Yelp, and TripAdvisor, with GMB having, at 11%, the highest percentage of fake reviews. [72] These results were based on reviews that had already passed through the platforms' own automated filters.

Automated detection efforts in this area are certainly worthwhile endeavors. As with other types of deceptive content, however, fake reviews remain hard to spot by their text alone or even via analysis of metadata. The fact that they remain a marketplace problem[73] indicates that current detection technology — even assuming sufficient investment therein by any given platform along with human oversight — is still not good enough.[74]

---

and detecting fake reviews in app stores, Empir. .Software Eng. 24: 3316–3355 (2019), https://doi.org/10.1007/s10664-019-09706-9; Arjun Mukherjee et al., *Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews* (2013), http://www2.cs.uh.edu/~arjun/tr/UIC-CS-TR-yelp-spam.pdf.

[69] *See* https://intercom help/fakespot/en/articles/2700070-analysis-criteria; https://reviewmeta.com/forum/index.php?/topic/173-why-not-use-a-machine-learning-algo-that-trains-itself-based-on-removed-reviews/.

[70] Fakespot, *US Online Shopping Ratings & Reviews Analysis Report* (2021), https://www.fakespot.com/2021holidayreport.

[71] *Id.*

[72] *See* Uberall, *The State of Online Review Fraud* (2021), https://uberall.com/en-us/resources/blog/how-big-a-problem-are-fake-reviews.

[73] In statements filed with the 2019 announcement of an action against Sunday Riley Modern Skincare, every FTC commissioner at that time recognized the serious harms caused by fake reviews. *See* https://www.ftc.gov/system/files/documents/cases/2020.11.6_sunday_riley_majority_statement_final.pdf and https://www.ftc.gov/system/files/documents/public_statements/1550127/192_3008_final_rc_statement_on_sunday_riley.pdf.

[74] *See, e.g.*, Department of Homeland Security, *Combating Trafficking in Counterfeit and Pirated Goods* (2020) ("the ratings systems across platforms have been gamed, and the proliferation of fake reviews and counterfeit goods on third-party marketplaces now threatens the trust mechanism itself"), https://www.dhs.gov/publication/combating-trafficking-counterfeit-and-pirated-goods; Competition and Markets Authority (United Kingdom), *Algorithms: How they can reduce competition and harm consumers* at 33-34 (2021), https://www.gov.uk/government/publications/algorithms-how-they-can-reduce-competition-and-harm-consumers/algorithms-how-they-can-reduce-competition-and-harm-consumers; CHEQ, *Fake Online Reviews* (2021), https://www.cheq.ai/research; Katie Schoolov, *Amazon is filled with fake reviews and it's getting harder to spot them*, CNBC (Sep. 6, 2000), https://www.cnbc.com/2020/09/06/amazon-reviews-thousands-are-fake-heres-how-to-spot-them html; George Nguyen, *How Google and Yelp handle fake reviews and policy violations*, Search Engine Land (Aug. 30, 2021), https://searchengineland.com/how-google-and-yelp-handle-fake-reviews-and-policy-violations-374071.

Fake accounts

Fake accounts on online platforms, often driven by bots, are themselves a form of manipulative content and serve the widely varying, manipulative intent of their operators. In 2020, the Commission reported to Congress on the use of social media bots in advertising, citing studies showing that, despite ongoing detection efforts, such bots remain hard to detect and easily capable of conducting widespread social media manipulation.[75]

Some platforms have been developing and using AI tools to detect such accounts and other inauthentic activity, including Facebook and its Instagram and WhatsApp properties.[76] Apple has indicated that it, too, uses machine learning to detect if users are real people.[77] TikTok reports that it uses automated tools to detect fake accounts and engagement.[78] However, as a State Department report, former FTC commissioner Rohit Chopra, and others have argued, social media platforms have strong financial incentives not to police this problem adequately.[79]

Often with federal funding, researchers have also been developing AI-based methods, including publicly available tools, to detect fake social media accounts and bots,[80] as well as state-

---

[75] *See* FTC Report to Congress on Social Media Bots and Advertising (Jul. 16, 2020), https://www.ftc.gov/reports/social-media-bots-advertising-ftc-report-congress. *See also* Sebastian Bay and Rolf Fredheim, *Social Media Manipulation 2021/2022: Assessing the Ability of Social Media Companies to Combat Platform Manipulation*, NATO Strategic Communications Centre of Excellence (Apr. 2022), https://stratcomcoe.org/publications/social-media-manipulation-20212022-assessing-the-ability-of-social-media-companies-to-combat-platform-manipulation/242.

[76] *See* Karen Hao, *How Facebook uses machine learning to detect fake accounts*, MIT Tech. Rev. (Mar. 4, 2020), https://www.technologyreview.com/2020/03/04/905551/how-facebook-uses-machine-learning-to-detect-fake-accounts/; https://research.fb.com/blog/2020/04/detecting-fake-accounts-on-social-networks-with-sybiledge/; https://business.instagram.com/blog/reducing-inauthentic-activity-on-instagram; https://faq.whatsapp.com/general/security-and-privacy/unauthorized-use-of-automated-or-bulk-messaging-on-whatsapp/?lang=en. In 2021, Facebook also disclosed that it demotes content associated with suspected fake accounts, such as instances of inauthentic sharing and posts from pages with artificially inflated distribution, though it does not indicate what tools it uses to identify such content. *See* https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/.

[77] *See* https://developer.apple.com/documentation/sign_in_with_apple/sign_in_with_apple_rest_api/authenticating_users_with_sign_in_with_apple/.

[78] *See* https://www.tiktok.com/safety/resources/tiktok-transparency-report-2021-q-2?lang=en.

[79] *See* Christina Nemr and William Gangware, *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*, Park Advisors at 26 (Mar. 2019), https://www.park-advisors.com/disinforeport; Rohit Chopra Statement, *Report to Congress on Social Media Bots and Deceptive Advertising* (Jul. 16, 2020), https://www.ftc.gov/public-statements/2020/07/statement-commissioner-rohit-chopra-regarding-report-congress-social-media; Simone Stolzoff, *The Problem with Social Media Has Never Been About Bots. It's Always Been About Business Models*, Quartz (Nov. 16, 2018), https://qz.com/1449402/how-to-solve-social-medias-bot-problem/.

[80] *See, e.g.*, Iacopo Pozzana and Emilio Ferrara, *Measuring Bot and Human Behavioral Dynamics*, Frontiers in Physics (Apr. 22, 2020) (citing the pioneering and extensive research in this area as well as openly accessible detection tools), https://www.frontiersin.org/articles/10.3389/fphy.2020.00125/full; Mohsen Sayyadiharikandeh, et al., *Detection of Novel Social Bots by Ensembles of Specialized Classifiers* (Aug. 14, 2020),

sponsored troll accounts.[81] Unfortunately, many of these tools are limited to Twitter because other platforms, like Facebook, restrict their APIs in ways that prevent access to the data necessary to create and test such tools.[82] The need to increase research access generally is discussed below.

As with deepfakes, one can expect the battle to continue between those seeking to detect fake accounts and those developing ever more sophisticated ways to deploy them for illicit purposes.

C. Website or mobile application interfaces designed to intentionally mislead or exploit individuals

This category of harm appears to refer principally to so-called "dark patterns," which were the focus of a 2021 Commission public workshop and a later Enforcement Policy Statement.[83] The potential use of AI to detect dark patterns has not been fully explored.[84] It may be that the creation of effective detection tools will remain challenging for the same reasons as noted above with respect to fraudulent and deceptive content generally. Another challenge is the need to resolve complex issues of how to define, identify, and measure dark patterns,[85] which would presumably be a precondition for setting computers to the same task. However, one oft-cited research study used automated tools to help detect dark patterns on shopping sites.[86] Further, the

---

https://arxiv.org/pdf/2006.06867.pdf; Adrian Rauchfleisch and Jonas Kaiser, *The False positive problem of automatic bot detection in social science research*, PLoS ONE 15(10): e0241045 (Oct. 22, 2020), https://doi.org/10.1371/journal.pone.0241045.

[81] *See* Mohammad Hammas Saeed, et al., *TROLLMAGNIFIER: Detecting State-Sponsored Troll Accounts on Reddit* (Dec. 1, 2021), https://arxiv.org/pdf/2112.00443.pdf; Chris Stokel-Walker, *Researchers Have a Method to Spot Reddit's State-Backed Trolls*, WIRED UK (Jan. 12, 2021), https://www.wired.co.uk/article/researchers-reddit-state-trolls.

[82] *See* EPRS, *Automated Tackling of Disinformation* at 33-34 (Mar. 2019), https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf; Johanna Wild and Charlotte Godart, *Spotting bots, cyborgs and inauthentic activity*, in Verification Handbook for Disinformation and Media Manipulation (Craig Silverman, ed.) (2020), https://datajournalism.com/read/handbook/verification-3.

[83] *See* https://www ftc.gov/news-events/events-calendar/bringing-dark-patterns-light-ftc-workshop; https://www.ftc.gov/news-events/press-releases/2021/10/ftc-ramp-enforcement-against-illegal-dark-patterns-trick-or-trap. *See also* Arvind Narayanan, et al., *Dark Patterns: Past, Present, and Future*, Queue (Mar.-Apr. 2020), https://dl.acm.org/doi/pdf/10.1145/3400899.3400901.

[84] *See* Competition and Markets Authority, *Online Choice Architecture: How digital design can harm competition and consumers* at 42 (Apr. 5, 2022), https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers.

[85] *See* Jennifer King and Adriana Stephan, *Regulating Privacy Dark Patterns in Practice — Drawing Inspiration from California Privacy Rights Act*, 5 Geo. L. Tech. Rev. 250 (2021), https://georgetownlawtechreview.org/wp-content/uploads/2021/09/King-Stephan-Dark-Patterns-5-GEO.-TECH.-REV.-251-2021.pdf. Among other things, it would be difficult to determine what training data one would use to build a dark pattern detection model.

[86] *See* Arunesh Mathur et al., *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*, Proc. of the ACM Human-Computer Interaction, Vol. 3, CSCW, Art. 81 (Nov. 2019), https://arxiv.org/abs/1907.07032. *See also*

German government is funding a project to create an AI-based app for detecting dark patterns.[87] Also worth noting is a project at Stanford University's Institute for Human-Centered AI, in which researchers are collecting and analyzing data on dark patterns and will then try to classify new ones in the wild.[88]

D. Illegal content online, including the illegal sale of opioids, child sexual exploitation and abuse, revenge pornography, harassment, cyberstalking, hate crimes, the glorification of violence or gore, and incitement of violence

Illegal sales of opioids and other drugs

Multiple federal agencies have been looking into developing AI tools as a way to detect illegal opioid sales or disrupt opioid traffickers. The National Institute on Drug Abuse, which is part of the Department of Health and Human Services, has invested in the creation of an AI-based tool to detect illegal opioid sellers.[89] The National Institute of Justice, which is part of the Department of Justice, has invested in AI technology to expose opioid trafficking on the dark web.[90] Further, the Food and Drug Administration has indicated that it uses AI-enabled tools in the context of its criminal investigations.[91]

Social media companies are reportedly using AI and other means to root out opioid and other illegal drug sales,[92] though such drugs can still easily be found for sale on those sites.[93] This

---

OECD, *Roundtable on Dark Commercial Patterns Online: Summary of discussion* at 6 (2021) (suggesting collaboration between consumer protection authorities and academics to develop automated detection tools), https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CP(2020)23/FINAL&docLanguage=En.

[87] *See* https://dapde.de/en/project/teilbereich-informatik-en/.

[88] *See* Katherine Miller, *Can't Unsubscribe? Blame Dark Patterns,* Stanford HAI News (Dec. 13, 2021), https://hai.stanford.edu/news/cant-unsubscribe-blame-dark-patterns.

[89] *See* https://www.usaspending.gov/award/CONT_AWD_75N95019C00069_7529_-NONE-_-NONE-.

[90] *See* https://nij.ojp.gov/funding/awards/2018-75-cx-0032.

[91] *See* Rebecca Heilweil, *AI can help find illegal opioid sellers online. And wildlife traffickers. And counterfeits*, Vox recode (Jan. 21, 2020), https://www.vox.com/recode/2020/1/21/21060680/opioids-artificial-intelligence-illegal-online-pharmacies.

[92] *See, e.g.*, https://snap.com/en-US/safety-and-impact/post/expanding-our-work-to-combat-the-fentanyl-epidemic (reporting that Snap also directs people searching for drug content to an educational portal and that it is constantly updating its databases to account for new drug terms that illicit drug sellers employ); https://transparency fb.com/data/community-standards-enforcement/regulated-goods/facebook/.

[93] *See, e.g.*, Jan Hoffman, *Fentanyl Tainted Pills Bought on Social Media Cause Youth Drug Deaths to Soar*, The New York Times (May 19, 2022), https://www nytimes.com/2022/05/19/health/pills-fentanyl-social-media.html; Tech Transparency Project, *Spot Check: Instagram's Drug Pipeline for Teens* (May 17, 2022), https://www.techtransparencyproject.org/articles/spot-check-instagrams-drug-pipeline-teens; Olivia Solon and Zoe Schiffer, *Instagram pushes drug content to teens*, NBC News (Dec. 7, 2021), https://www.nbcnews.com/tech/social-

situation reflects the huge amount of content to be policed, the fact that drug dealers keep devising sophisticated methods to trick the detection algorithms, and the need for more research into and constant improvement of such detection methods.[94]

Professor Tim Mackey of the University of California, San Diego, has led several government-funded efforts in this area and has published studies on the use of AI to detect illegal sales of online drugs and COVID-19 health products.[95] The tools developed from this work could potentially be used to track drug sales by location, help law enforcement link online and offline investigations, reveal elements of the supply chain, and perhaps redirect those seeking opioids to rehabilitative resources.[96]

Child sex exploitation and abuse

Several major technology companies collaborate to address child sexual abuse material (CSAM) via the Technology Coalition, which publishes annual reports on industry efforts.[97] These companies use automated tools, including a hash-matching[98] technology from Microsoft called PhotoDNA, to identify and remove CSAM.[99] This process involves organizations like the National Center for Missing & Exploited Children assigning unique, "hash-based" alphanumeric identifiers to images of known CSAM; platforms then compile and use those hashes — which use a common format across industry — to block attempts to upload known CSAM.[100] According to the Technology Coalition, hash-based *video* detection is "less developed," with fewer members using such tools and without "an industry standard hash format."[101] Other joint efforts include the WeProtect Global Alliance, a public-private collaboration that, among other

media/instagram-pushes-drug-content-teens-rcna7751; Olivia Solon, *Snapchat boosts efforts to root out drug dealers*, NBC News (Oct. 7, 2021), https://www.nbcnews.com/tech/social-media/snapchat-boosts-efforts-root-out-drug-dealers-n1280946?s=03; Rachel Lerman and Gerrit De Vynck, *Snapchat, TikTok, Instagram face pressure to stop illegal drug sales as overdose deaths soar*, The Washington Post (Sep. 28, 2021), https://www.washingtonpost.com/technology/2021/09/28/tiktok-snapchat-fentanyl/; Heilweil, *supra* note 91.

[94] *Id.*

[95] *See, e.g.*, Neal Shah, et al., *An unsupervised machine learning approach for the detection and characterization of illicit drug-dealing comments and interactions on Instagram,* Substance Abuse, https://www.tandfonline.com/doi/abs/10.1080/08897077.2021.1941508; Tim Mackey et al., *Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales*, JMIR Public Health Surveill. 6(3): e20794 (Jul.-Sep. 2020), https://publichealth.jmir.org/2020/3/e20794/; Tim Mackey, et al., *Twitter-Based Detection of Illegal Online Sale of Prescription Opioid*, Am J Public Health 107(12): 1910–1915 (Dec. 2017), https://ajph.aphapublications.org/doi/10.2105/AJPH.2017.303994.

[96] Heilweil, *supra* note 91.

[97] *See* Technology Coalition Annual Report 2021, https://technologycoalition.org/annualreport/.

[98] *See generally* Hany Farid, *An Overview of Perceptual Hashing*, J. Online Trust and Safety (Oct. 2021), https://tsjournal.org/index.php/jots/article/view/24/14.

[99] *See id.*

[100] *See id.* A Canadian effort, Project Arachnid, also uses matching tools. *See* https://projectarachnid.ca/en/#how-does-it-work.

[101] *See* Technology Coalition, *supra* note 97.

things, surveys companies about their detection efforts and makes recommendations.[102] Thorn, a nonprofit entity, offers a hash-matching tool, Safer, to content-hosting sites.[103]

Hash-matching is not AI, but some companies have developed AI tools as a way to flag new or unhashed CSAM. The Technology Coalition reports that its members use a variety of classifiers – algorithms supported by machine learning — to flag potential CSAM for categorization and human review.[104] These classifiers, which are often open source, include Google's Content Safety API.[105] Facebook also uses AI tools to spot new or unhashed CSAM,[106] and some service providers offer such tools to platforms.[107] Law enforcement around the world also uses third-party AI tools to detect and evaluate CSAM in videos or images.[108]

Separately, in 2021, Apple announced that it will provide an opt-in setting in family iCloud accounts that uses on-device machine learning to detect sexually explicit photos sent in the Messages app.[109] The system can display warnings to children when such photos are being sent or received, but Apple will not get access to the messages.[110] Apple decided to delay rollout of other announced measures to deal with CSAM when security and privacy experts raised concerns about potential misuse of new device-scanning technology.[111]

---

[102] *See* https://www.weprotect.org/.

[103] *See* https://www.thorn.org/. *See also* Caroline Donnelly, *Thorn CEO on using machine learning and tech partnerships to tackle online child sex abuse*, Computer Weekly (Mar. 29, 2017), https://www.computerweekly.com/news/450415609/Thorn-CEO-on-using-machine-learning-and-tech-partnerships-to-tackle-online-child-sex-abuse.

[104] *See* Technology Coalition, *supra* note 97.

[105] *See id.*; https://protectingchildren.google/intl/en/#tools-to-fight-csam.

[106] *See* https://about fb.com/news/2018/10/fighting-child-exploitation/.

[107] *See, e.g.*, https://www.twohat.com/cease-ai/.

[108] *See, e.g.*, https://www.griffeye.com/griffeye-releases-new-ai-that-can-identify-csa-content-in-videos/#; https://news.microsoft.com/de-de/ki-im-einsatz-gegen-kinderpornografie/; Matt Burgess, *AI is helping UK police tackle child abuse way quicker than before*, WIRED UK (Jul. 17, 2019), https://www.wired.co.uk/article/uk-police-child-abuse-images-ai; Anouk Vleugels, *AI algorithms identify pedophiles for the police — here's how it works*, The Next Web. (Nov. 8, 2018), https://thenextweb.com/news/ai-algorithms-identify-sexual-child-abuse-for-the-police.

[109] *See* https://www.apple.com/child-safety/.

[110] *See id.*

[111] *See* Reed Albergotti, *Apple delays the rollout of its plans to scan iPhones for child exploitation images*, The Washington Post (Sep. 3, 2021), https://www.washingtonpost.com/technology/2021/09/03/apple-delay-csam-scanning/; Jonathan Mayer and Anunay Kulshrestha, *Opinion: We built a system like Apple's to flag child sexual abuse material — and concluded the tech was dangerous*, The Washington Post (Aug. 19, 2021), https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/; Hany Farid, *Opinion: Should we Celebrate or Condemn Apple's New Child Protection Measures?*, Newsweek (Aug. 13, 2021), https://www newsweek.com/should-we-celebrate-condemn-apples-new-child-protection-measures-opinion-1618828?amp=1&__twitter_impression=true. *See also* Nat Rubio-Licht, *Apple will soon blur nude photos sent to kids' iPhones*, Protocol (Apr. 20, 2022) (Apple using blue feature only in UK for messages with nude images sent to or from children), https://www.protocol.com/apple-message-scan-csam.

Other platform-developed tools deal with the related problem of child grooming.[112] Instagram uses AI tools that prevent adults from sending messages to people under 18 who don't follow them, sends prompts or safety notices to encourage teens to be cautious in conversations with adults to whom they are already connected but who are exhibiting potentially suspicious behavior, and prevent such adults from interacting with teens.[113] A Microsoft tool, Project Artemis, uses machine learning to detect child grooming by reviewing chat features of video games and messaging apps for patterns of communication that predators use to target children; the tool flags that content for human reviewers who decide whether to contact law enforcement.[114]

The research community is also studying CSAM detection methods with the help of AI. One study synthesized this work and concluded that the best results may occur when detection approaches are used in combination, and that deep learning techniques outperform other methods for detecting unknown CSAM.[115] Other researchers are taking different paths, such as the H-Unique project, centered at the United Kingdom's Lancaster University, involving an interdisciplinary study of the anatomical differences of hands.[116] If all hands are truly unique, then computers can be trained to identify someone's hand from a photograph, and algorithms can be designed to link those images to crime suspects.[117] That's especially important for certain child sexual abuse cases, where the only visible features of the abuser may be the backs of their hands seen in photographs.[118]

Detection of this kind of material is obviously important, and development of appropriate and effective tools should continue.[119] As reflected by the examples above, some platforms are actively engaged, taking usually positive though sometimes controversial measures. Other platforms and industry in general have been criticized for moving slowly and unevenly, not using

---

[112] Grooming involves a predator or pornographer fostering a false sense of trust and authority over a child in order to desensitize or break down the child´s resistance to sexual abuse. *See* https://www.justice.gov/criminal-ceos/child-pornography.

[113] *See* https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community.

[114] *See* https://blogs.microsoft.com/on-the-issues/2020/01/09/artemis-online-grooming-detection/.

[115] *See* Hee-Eun Lee, et al., *Detecting child sexual abuse material: A comprehensive survey*, Forensic Science International: Digital Investigation 34 (Sep. 2020), https://doi.org/10.1016/j fsidi.2020.301022. *See also* Elie Burzstein et al., *Rethinking the Detection of Child Sexual Abuse Imagery on the Internet*, in Proc. of the 2019 World Wide Web Conference (May 2019), https://doi.org/10.1145/3308558.3313482.

[116] *See* https://www.lancaster.ac.uk/security-lancaster/research/h-unique/.

[117] *See* https://www.lancaster.ac.uk/news/new-app-launched-for-public-to-help-pioneering-hand-identification-research#:~:text=Led%20by%20forensic%20anthropologist%20Professor,the%20environment%20and%20even%20accidents.

[118] *See id.*

[119] Indeed, CSAM may present the case where automated detection is clearly the most useful strategy for detection. Riana Pfefferkorn, *Content-Oblivious Trust and Safety Techniques: Results from a Survey of Online Service Providers* (Sep. 9, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3920031.

all tools at their disposal, and not being transparent.[120] A 2019 New York Times report describes flaws in search engine filtering of such material and notes that Amazon Web Services does not search for CSAM at all, which appears still to be the case today.[121] More recently, WeProtect issued an annual report expressing hope but highlighting the growing scale of this material online and the need for, among other things, continued technological innovation and collaboration.[122] The extensive report describes the state of detection efforts, technological limits and problems (such as end-to-end encryption), and the failure of some platforms to use available tools.[123]

Revenge pornography

Automated detection of revenge pornography — the nonconsensual sharing of intimate images — has not received much attention, at least not relative to many other categories of harm discussed here.[124] This fact may reflect the difficulty of training a machine to determine the nonconsensual nature of an image or video — a determination that humans, too, may not always be able to make easily. The need for such determinations also distinguishes this category from CSAM, where context and intent are not at issue. Nonetheless, Facebook has invested in creating an AI tool for this purpose, one that looks at patterns in the language accompanying an image, [125] as well as programs involving reporting by victim advocates and digital fingerprinting of images to prevent malicious upload.[126] We are unaware of whether other platforms or researchers have

---

[120] *See* Internet Watch Foundation, *The Annual Report 2021* at 14 (statement of Hany Farid), https://www.iwf.org.uk/about-us/who-we-are/annual-report-2021/; Michael H. Keller and Gabriel J.X. Dance, *Child Abusers Run Rampant as Tech Companies Look the Other Way*, New York Times (Nov. 9, 2019), https://www.nytimes.com/interactive/2019/11/09/us/internet-child-sex-abuse html.

[121] *See id.*; Sheila Dang, *Amazon considers more proactive approach to determining what belongs on its cloud service*, Reuters (Sep. 5, 2021) (quoting an AWS spokesperson that it "does not pre-review content hosted by our customers" and stating that it has no intent to scan existing content), https://www reuters.com/technology/exclusive-amazon-proactively-remove-more-content-that-violates-rules-cloud-2021-09-02/.

[122] *See* WeProtect Global Alliance, *Global Threat Assessment 2021*, https://www.weprotect.org/global-threat-assessment-21/. See also Internet Watch Foundation, *supra* note 120 at 99.

[123] *Id. See also* Broadband Commission for Sustainable Development, *Child Online Safety: Minimizing the Risk of Violence, Abuse and Exploitation Online* 37-38 (Oct. 2019), https://broadbandcommission.org/wp-content/uploads/2021/02/ChildOnlineSafety_Report.pdf.

[124] The FTC has brought actions against companies involved in posting such images and charging takedown fees. *See* https://www ftc.gov/news-events/press-releases/2018/06/ftc-nevada-obtain-order-permanently-shutting-down-revenge-porn; https://www.ftc.gov/news-events/press-releases/2016/01/ftc-approves-final-order-craig-brittain-revenge-porn-case.

[125] *See* https://about fb.com/news/2019/03/detecting-non-consensual-intimate-images/; Nicola Henry and Alice Witt, *Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, and Practices*, in The Emerald International Handbook of Technology-Facilitated Violence and Abuse at 758-59 (Jun. 4, 2021), https://doi.org/10.1108/978-1-83982-848-520211054; Solon, *Inside Facebook's efforts to stop revenge porn before it spreads, supra* note 8.

[126] *See id.*; https://www facebook.com/safety/notwithoutmyconsent/pilot/how-it-works; Danielle Keats Citron, *Sexual Privacy*, 128 Yale L.J. 1870, 1955-58 (2019), https://digitalcommons.law.yale.edu/ylj/vol128/iss7/2/. It appears that one of these programs was dropped for unknown reasons, *see* Elizabeth Dwoskin and Craig Timberg, *Like whistleblower Frances Haugen, these Facebook employees warned about the company's problems for years.*

engaged in similar work to date, although this harm is often connected with deepfakes, discussed above.

Hate crimes

As a preliminary matter, we note that Congress lists *hate crimes* as a form of illegal content on which this report should focus but does not include the related category of *hate speech*. Whereas hate crimes refer to criminal offenses intentionally directed at specific individuals, hate speech generally refers to communications about groups or classes of people.[127] This omission likely reflects the fact that, while harmful, hate speech is not illegal unless it amounts to threats or incitement to commit crimes.[128] Its legal status notwithstanding, the spread of online hate and the extent to which AI or other sophisticated technology can address it is the subject of much controversy and research.[129] Less explored is the question of whether such tools can detect or otherwise address hate crimes specifically. As automated tools are generally not proficient at detecting a hard-to-define and context-dependent category like hate speech,[130] though, it is hard

---

No one listened, The Washington Post (Oct. 8, 2021), https://www.washingtonpost.com/technology/2021/10/08/facebook-whistleblowers-public-integrity-haugen/, but that another one survived, *see* Olivia Solon, *Meta builds tool to stop the spread of 'revenge porn,'* NBC News (Dec. 2, 2021), https://www nbcnews.com/tech/social-media/meta-builds-tool-stop-spread-revenge-porn-rcna7231.

[127] *See* https://www.justice.gov/hatecrimes/learn-about-hate-crimes/chart; Department of Justice, *Investigating Hate Crimes on the Internet* (2003), https://www.ojp.gov/ncjrs/virtual-library/abstracts/investigating-hate-crimes-internet; United Nations Strategy and Plan of Action on Hate Speech at 2 (Jun. 2019), https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

[128] *See id.*

[129] *See, e.g.*, Deepa Seetharaman, et al., *Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts*, Wall St. J. (Oct. 17, 2021) (discussing small percentages of hate speech caught by platform using automated tools), https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184; Paul Rottger, et al., *HATECHECK: Functional Tests for Hate Speech Detection Models*, (May 27, 2021) (revealing critical weaknesses in detection models including Google Jigsaw's Perspective), https://arxiv.org/abs/2012.15606.

[130] Even putting aside technical limits, a foundational problem is that hate speech is not easily definable, or at least is not amenable to easy agreement on its definition, making it even more difficult to deploy effective detection tools. *See, e.g.*, Adam G. Klein, *Fear, more than hate, feeds online bigotry and real-world violence*, The Conversation, (Dec. 20, 2018), https://theconversation.com/fear-more-than-hate-feeds-online-bigotry-and-real-world-violence-106988; EPRS, *Automated tackling of disinformation*, *supra* note 82 at 39. Similarly, given the difficult contextual judgments required, which involve sensitivity to different cultures and languages, humans and machines can both fail easily when trying to determine if certain posts fit a definition. *See, e.g.*, Facebook Oversight Board, *Case decision 2021-007-FB-UA*, https://www.oversightboard.com/decision/FB-ZWQUPZLZ; Tekla S. Perry, *Q&A: Facebook's CTO Is at War With Bad Content, and AI Is His Best Weapon*, IEEE Spectrum (Jul. 21, 2020) (Mike Schroepfer noting how language and context make it hard to use AI to detect hate speech), https://spectrum.ieee.org/computing/software/qa-facebooks-cto-is-at-war-with-bad-content-and-ai-is-his-best-weapon; Jennifer Young, et al., *Beyond AI: Responses to Hate Speech and Disinformation*, Carnegie Mellon U. (2018), https://jessica-young.com/research/Beyond-AI-Responses-to-Hate-Speech-and-Disinformation.pdf. Definitional and contextual issues are discussed more in Section IV.

to conceive that such tools could easily distinguish when given hateful content is more or less likely to be criminal.[131]

On the other hand, while AI tools might not be good enough at detecting hateful content,[132] they might help in other ways, such as by predicting when hate speech may lead to violence and crime in the physical world.[133] At least three sets of academics have probed such correlations:

- A New York University research team, with partial federal funding, used machine learning to show that cities with a greater incidence of a certain type of racist post on Twitter reported more hate crimes related to race, ethnicity, and national origin.[134]

- Researchers from Cardiff University's Hatelab project collected Twitter data via an AI tool and compared it to London police data to show that an increase in "hate tweets" from one location corresponded to an increase in racially and religiously aggravated crimes in the same area.[135] The Cardiff researchers, supported in part by the United States Department of Justice, suggested that an algorithm using their method could predict spikes in crimes against members of minority communities in specific areas.[136]

- Researchers from Princeton University and the University of Warwick, using methods including machine learning, found correlations between increases in Twitter usage and anti-Muslim hate crimes in certain United States counties since the 2016 Presidential election.[137] In a separate study, also using a machine learning tool and focused on Germany, they determined that "anti-

---

[131] It is worth noting that hate crime has been a vexed area for enforcement, with statistics indicating that, while hate crimes against racial minorities are under-reported, hate crime laws are enforced disproportionately against those same minorities. *See, e.g.*, Stanford Law School Policy Lab and Brennan Center for Justice, *Exploring Alternative Approaches to Hate Crimes* at 13-14 (Jun. 2021), https://www-cdn.law.stanford.edu/wp-content/uploads/2021/06/Alternative-to-Hate-Crimes-Report_v09-final.pdf; Michael German and Emmanuel Mauleón, *Fighting Far Right Violence and Hate Crimes* at 14, Brennan Center for Justice (Jul. 1, 2019), https://www.brennancenter.org/sites/default/files/2019-08/Report_Far_Right_Violence.pdf; Heather Zaykowski, *Racial Disparities in Hate Crime Reporting*, Violence and Victims 25:3 (Jun. 2010), https://doi.org/10.1891/0886-6708.25.3.378. AI detection systems and platform policies that rely on historical crime data may thus be likely to reflect these disparities.

[132] Of course, the limitations of automated approaches should not diminish continued work in this area, such as the positive efforts of the Anti-Defamation League, *see* https://www.adl.org/resources/reports/the-online-hate-index, and the Alan Turing Institute, *see* https://www.turing.ac.uk/blog/introducing-online-harms-observatory.

[133] *See generally* Cathy Buerger, *Speech as a Driver of Intergroup Violence: A Literature Review*, Dangerous Speech Project (Jun. 16, 2021), https://dangerousspeech.org/wp-content/uploads/2021/06/Speech-and-Violence-Lit-Review.pdf.

[134] *See* Kunal Relia et al., *Race, Ethnicity and National Origin-based Discrimination in Social Media and Hate Crimes Across 100 U.S. Cities* (Jan. 31, 2019), https://arxiv.org/pdf/1902.00119.pdf.

[135] *See* Matthew L. Williams, et al., *Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime*, Brit. J. Criminol. 60, 93–117 (Jul. 23, 2019), https://doi.org/10.1093/bjc/azz049.

[136] *See id.*

[137] Karsten Muller and Carlo Schwarz, *From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment* (Jul. 24, 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103.

> refugee sentiment on Facebook predicts crimes against refugees in otherwise similar municipalities with higher social media usage."[138]

Employees of at least one social media platform have focused on this link, too. Internal Facebook documents show that analysts worried that hateful content on the platform might be inciting real-world violence in connection with Minneapolis protests occurring after the police killing of George Floyd.[139] Although it is not clear what precise tools they used, these analysts discovered that "the largest and most combative demonstrations" took place in two zip codes where users reported spikes in offensive posts, whereas harmful content was only "sporadic" in areas where protests had not yet emerged.[140]

Harassment and cyberstalking

The Commission has brought multiple cases against stalkerware app companies.[141] AI tools could aid in detecting similar apps. Researchers at Cornell and New York University worked with NortonLifeLock to create CreepRank, an algorithm that ranks the probability that an app is used as "creepware" — hard-to-detect software that can be used to abuse, stalk, harass and spy on others.[142] NortonLifeLock incorporated it into its mobile security service, and the researchers reported suspect apps to Google, which removed over 800 of them from the Play Store.[143] The study did not use AI, but the researchers note that CreepRank could be a first step in collecting and using data that would train machine learning classifiers to identify these apps.[144]

Building automated tools to detect particular incidents of harassment or cyberstalking is challenging for the same reasons as described above with respect to hate crimes. Professor Citron has noted, both in her seminal work, Hate Crimes in Cyberspace, and thereafter, that, in connection with harassment and threats, computers cannot yet approximate the contextual judgment of humans.[145] A recent Google Research paper delves into this and other challenges of

---

[138] Karsten Muller and Carlo Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime* (Jun. 8, 2020), https://ssrn.com/abstract=3082972.

[139] *See* Naomi Nix and Lauren Etter*, Facebook Privately Worried About Hate Speech Spawning Violence*, Bloomberg.com (Oct. 25, 2021), https://www.bloomberg.com/news/articles/2021-10-25/facebook-s-fb-hate-speech-problem-worried-its-own-analysts.

[140] *Id.*

[141] *See, e.g.*, https://www ftc.gov/news-events/press-releases/2021/09/ftc-bans-spyfone-and-ceo-from-surveillance-business; https://www ftc.gov/news-events/press-releases/2020/03/ftc-gives-final-approval-settlement-stalking-apps-developer.

[142] *See* https://www.nortonlifelock.com/blogs/research-group/what-were-doing-fight-scourge-cyber-stalking.

[143] *See* Kevin A. Roundy, et al., *The Many Kinds of Creepware Used for Interpersonal Attacks*, 2020 IEEE Symposium on Security and Privacy (2020) https://ieeexplore.ieee.org/ielx7/9144328/9152199/09152794.pdf.

[144] *Id. See also* Ingo Frommholz, et al., *On Textual Analysis and Machine Learning for Cyberstalking Detection*, Datenbank Spektrum 16:127–135 (2016), https://link.springer.com/article/10.1007/s13222-016-0221-x.

[145] *See* Danielle Keats Citron, Hate Crimes in Cyberspace at 232 (2014); Danielle Keats Citron, *Section 230's Challenge to Civil Rights and Civil Liberties*, Knight First Amendment Institute, at n.41 (Apr. 6, 2018), https://knightcolumbia.org/content/section-230s-challenge-civil-rights-and-civil-liberties. *See also* Erik Larson, The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do (2021).

automating detection of hate and harassment; it reviewed past studies and noted that classifiers can be designed not simply to detect individual instances but also to identify abusive accounts or predict at-risk users, and that "classifier scores can feed into moderation queues, content ranking algorithms, or warnings and nudges."[146] These researchers — and others before them — have explained, however, that all of these strategies struggle with obtaining unbiased and representative datasets of abusive content for training.[147]

Nonetheless, companies have focused some AI-related efforts in at least one closely related area: cyberbullying.[148] For example, IBM has worked with several start-ups and the Megan Meier Foundation on tools that use AI to detect possible child bullying and to find it in social media.[149] Further, in 2019, Instagram began rolling out AI-powered features intended to limit bullying by notifying people before they post comments or captions that may be considered offensive.[150] YouTube and TikTok indicate that they use automation of some kind to detect and remove videos featuring harassment or bullying.[151] Microsoft uses AI-powered content moderation on its Xbox gaming platform to detect cyberbullying and violent threats, among other things.[152]

Current cyberbullying research includes work from the Socio-Technical Interaction Research Lab, led by Dr. Pamela Wisniewski, including projects on detecting cyberbullying and other online sexual risks based on a human-centered approach to the use of AI.[153] One of

---

[146] Kurt Thomas, et al., *SoK: Hate, Harassment, and the Changing Landscape of Online Abuse* at 12, Google Research (2021), https://research.google/pubs/pub49786/.

[147] *Id*. at 12 (noting that biased training data can result in classifiers that consider terms like "gay" and "black" as themselves reflecting hate or harassment); Lindsay Blackwell, et al., *Classification and Its Consequences for Online Harassment: Design Insights from HeartMob,* Proc. of the ACM on Human-Computer Interaction (Dec. 2017) (discussing promise and limits of AI-based detection and how classification of harassment can invalidate the harassment experiences of marginalized people whose experiences aren't considered typical as defined per the morals and values of those creating the classification system), https://www.researchgate net/publication/321636042. *See also* Rhiannon Williams, *Google is failing to enforce its own ban on ads for stalkerware*, MIT Tech. Rev. (May 12, 2022) (referring to failure of algorithms to stop ads for stalkerware), https://www.technologyreview.com/2022/05/12/1052125/google-failing-stalkerware-apps-ads-ban/.

[148] *See generally* Sameer Hinduja, *How Machine Learning Can Help Us Combat Online Abuse: A Primer*, The Cyberbullying Resource Center (2017), https://cyberbullying.org/machine-learning-can-help-us-combat-online-abuse-primer.

[149] *See* https://www.ibm.com/cloud/blog/ibm-cloud-services-working-together-for-competitive-advantage; https://www.identityguard.com/news/can-ai-solve-cyberbullying.

[150] *See* https://about.instagram.com/blog/announcements/our-progress-on-leading-the-fight-against-online-bullying.

[151] *See* https://transparencyreport.google.com/youtube-policy/removals; https://www.tiktok.com/safety/resources/tiktok-transparency-report-2021-q-2?lang=en.

[152] *See* Tom Warren, *Microsoft acquires Two Hat, a moderation company that helps keep Xbox clean*, The Verge (Oct. 29, 2021), https://www.theverge.com/2021/10/29/22752421/microsoft-two-hat-acquisition-xbox-moderation?scrolla=5eb6d68b7fedc32c19ef33b4&s=03; https://www.twohat.com/solutions/content-moderation-platform/.

[153] *See* https://stirlab.org/; Seunghyun Kim, et al., *A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms*, Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 325 (Oct. 2021), https://doi.org/10.1145/3476066; Afsaneh Razi, et al., *A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection*, Proc. ACM Hum.-Comput. Interact., Vol. 5, No. CSCW2, Article 465 (Oct. 2021), https://doi.org/10.1145/3479609.

Dr. Wisniewski's research projects also led to the creation of MOSafely.org, an open-source community that leverages AI, evidence, and data to address these online safety issues, supported by a federal grant.[154] Other work includes an EU project called Creep that uses AI to spot cyberbullying and distinguish it from simple disagreement, and that aims to develop prevention techniques via a chatbot.[155] An effort at the University of Exeter's Business School involves development of a tool, LOLA, that uses natural language processing to detect emotional undertones that may indicate cyberbullying.[156] Other researchers, sometimes with public funding, have used varying AI techniques to develop other detection methods.[157] Unsurprisingly, some researchers have raised the same problems with representative datasets, classifications, and definitions noted above.[158]

Glorification or incitement of violence

Many major tech platforms and companies have developed and use AI tools to attempt to filter different kinds of violent content.[159] For example, YouTube built classifiers in 2011 to identify violent videos and prevent them from being recommended.[160] That platform and TikTok have both indicated more recently that they use automated measures to detect and remove violent and graphic content.[161] Facebook also uses such tools,[162] as does Pinterest.[163] Further, Parler uses a content moderation platform operated by a third party, Hive, which, among other things,

---

[154] *See* https://www.mosafely.org/mission-statement/.

[155] *See* http://creep-project.eu/.

[156] *See* https://business-school.exeter.ac.uk/newsandevents/news/articles/emotiondetectionenginedev.html.

[157] *See, e.g.*, Jacopo De Angelis and Giulia Perasso, *Cyberbullying Detection Through Machine Learning: Can Technology Help to Prevent Internet Bullying?*, Int'l J. Mgmt. and Humanities 4(11) (Jul. 2020), https://www.ijmh.org/wp-content/uploads/papers/v4i11/K10560741120.pdf; Cynthia Van Hee, et al., *Automatic detection of cyberbullying in social media text*, PLoS One 13(10) (Oct. 8, 2018), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203794; Despoina Chatzakou, et al., *Mean Birds: Detecting Aggression and Bullying on Twitter* (May 12, 2017), https://arxiv.org/pdf/1702.06877.pdf.

[158] *See, e.g.*, Chris Emmery, et al., *Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity*, Lang. Resources & Eval. (2021) 55:597–633, https://doi.org/10.1007/s10579-020-09509-1; H. Rosa, et al., *Automatic cyberbullying detection: A systematic review*, Computers in Human Behavior, 93 (2019) 333-345, http://rosta-farzan.net/courses/SC2019/readings/Rosa2018.pdf.

[159] This subsection excludes the other specified harms that involve particular types of violence, which are discussed either above (hate crimes) or below (violent extremist content).

[160] *See* https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

[161] *See* https://transparencyreport.google.com/youtube-policy/removals; https://www.tiktok.com/safety/resources/tiktok-transparency-report-2021-q-2?lang=en.

[162] *See, e.g.*, https://ai facebook.com/blog/how-ai-is-learning-to-see-the-bigger-picture/; Dan Sabbagh, *Facebook trained its AI to block violent live streams after Christchurch attacks*, The Guardian (Oct. 29, 2021), https://www.theguardian.com/technology/2021/oct/29/facebook-trained-its-ai-to-block-violent-live-streams-after-christchurch-attacks.

[163] *See* Vishwakarma Singh and Dan Lee, *How Pinterest fights misinformation, hate speech, and self-harm content with machine learning*, Pinterest Engineering Blog (Mar. 5, 2021), https://medium.com/pinterest-engineering/how-pinterest-fights-misinformation-hate-speech-and-self-harm-content-with-machine-learning-1806b73b40ef.

removes content appearing to involve violence.[164] Amazon offers its Rekognition APIs to businesses for content moderation, including automated detection of violence and gore.[165]

It is generally unclear whether or to what extent these tools are effective in practice, given the lack of transparency about their use. In Facebook's case, however, leaked internal documents are not encouraging and contrast with its public representations. The Wall Street Journal reported that, in March 2021, a team of Facebook employees found that the company's automated systems removed only "0.6% of all content that violated Facebook's policies against violence and incitement."[166] An internal presentation from April 2020, focusing on prevalence instead of the total amount of content, found that "removals were reducing the overall prevalence of graphic violence by about 19 percent."[167]

One reason for skepticism about the use of AI for accurate detection of violent content is the familiar problem of context, noted already above and explored more below.[168] Nonetheless, worthwhile and varied research on violence detection methods has continued in the academic community, including, for example, a study in Mexico on using deep neural networks for gender-based violence detection in Twitter messages,[169] and development by Notre Dame researchers, with government funding, of an AI "early warning system" for manipulated media that may lead to violence.[170]

Unsafe or illegal items for sale

It does not appear that companies or researchers have done substantial work yet to develop AI tools to tackle this harm — from which we exclude the more specific categories of the illegal sale of drugs (discussed above) and the sale of counterfeit goods (discussed below). One

---

[164] *See* Kevin Randall, *Social app Parler is cracking down on hate speech — but only on iPhones*, The Washington Post (May 17, 2021), https://www.washingtonpost.com/technology/2021/05/17/parler-apple-app-store/; https://thehive.ai/.

[165] *See* https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html.

[166] Seetharaman, *supra* note 129. *See also* Olivia Little, *A network of TikTok accounts is teaching users how to make pipe bombs and other weapons*, Media Matters for America (May 17, 2022), https://www.mediamatters.org/tiktok/network-tiktok-accounts-teaching-users-how-make-pipe-bombs-and-other-weapons.

[167] Gilad Edelman, *How to Fix Facebook, According to Facebook Employees*, WIRED (Oct. 25, 2021), https://www.wired.com/story/how-to-fix-facebook-according-to-facebook-employees/?s=03.

[168] *See* Desmond U. Patton, et al., *Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing*, Proc. of the 2020 AAAI/ACM Conf. on AI, Ethics, and Society (Feb. 7-8, 2020), https://doi.org/10.1145/3375627.3375841; Rachel Metz, *Why AI is still terrible at spotting violence online*, CNN (Mar. 18, 2019) (explaining contextual problem of AI identifying incitement of violence in speech or violent imagery in video), https://www.cnn.com/2019/03/16/tech/ai-video-spotting-terror-violence-new-zealand/index html.

[169] Carlos M. Castorena, et al., *Deep Neural Network for Gender-Based Violence Detection on Twitter Messages*, Mathematics 9(8), 807 (2021), https://doi.org/10.3390/math9080807.

[170] Michael Yankoski, et al., *An AI early warning system to monitor online disinformation, stop violence, and protect elections*, Bulletin of the Atomic Scientists 76(2), 85-90 (2020), https://doi.org/10.1080/00963402.2020.1728976.

exception is Amazon, which developed machine learning tools to detect the sale of banned or unsafe goods in its marketplace, though the Wall Street Journal reported those measures have been ineffective.[171] Some researchers have used AI to detect online sales of particular items, such as illegal wildlife products sold on social media.[172] Another study used AI to detect likely food recalls and predict potentially unsafe food products based on analyses of Amazon customer reviews.[173]

E. Terrorist and violent extremists' abuse of digital platforms, including the use of such platforms to promote themselves, share propaganda, and glorify real-world acts of violence

DHS and others have recognized the importance of innovative technology in countering the online spread of terrorist and violent extremist content (TVEC). As early as 2017, a DHS advisory committee explained that AI systems "can be deployed in the counter-terror and countering violent extremism arenas to provide improvements to DHS capabilities."[174] In 2021, a DHS official described the agency's consideration of using companies — some of which employ AI tools — to find warning signs of extremist violence on social media.[175] As part of its CP3 initiative, DHS also announced the opening of the National Counterterrorism Innovation, Technology, and Education Center (NCITE), centered at the University of Nebraska.[176] Per a federal grant, NCITE researchers are attempting to create an intelligent chatbot that will improve

---

[171] *See* Alexandra Berzon, et al., *Amazon Has Ceded Control of Its Site. The Result: Thousands of Banned, Unsafe or Mislabeled Products*, Wall St. J. (Aug. 23, 2019), https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990. *See also* Melissa Heikkilä, *Online marketplaces rife with unsafe and illegal items, study shows*, Politico EU (Feb. 24, 2020), https://www.politico.eu/article/online-marketplaces-rife-with-unsafe-and-illegal-items-study-shows/; https://www.aboutamazon.com/news/company-news/product-safety-and-compliance-in-our-store.

[172] *See* Enrico Di Minin and Christoph Fink, *How machine learning can help fight illegal wildlife trade on social media*, The Conversation (Apr. 23, 2019), https://theconversation.com/how-machine-learning-can-help-fight-illegal-wildlife-trade-on-social-media-115021. *See also* Julio Hernandez-Castro and David L. Roberts, *Automatic detection of potentially illegal online sales of elephant ivory via data mining*, PeerJ Comput. Sci. 1:e10 (Jul. 2015), https://peerj.com/articles/cs-10/.

[173] Adyasha Maharana, et al., *Detecting reports of unsafe foods in consumer product reviews*, JAMIA Open 2(3), 330–338 (Oct. 2019), https://academic.oup.com/jamiaopen/article/2/3/330/5543660.

[174] Homeland Security Sci. and Techn. Advis. Comm., *Artificial Intelligence White Paper* (Mar. 10, 2017), https://www.dhs.gov/sites/default/files/publications/Artificial%20Intelligence%20Whitepaper%202017_508%20FINAL_2.pdf. *See also* Jonathan Fischbach, *A New AI Strategy to Combat Domestic Terrorism and Violent Extremism*, Harv. Nat'l Sec. J. Online (May 6, 2020) (discussing need for national security community to reassess effective use of AI in this area), https://harvardnsj.org/wp-content/uploads/sites/13/2020/05/Fischbach_A-New-AI-Strategy.pdf.

[175] *See* Rachael Levy, *Homeland Security Considers Outside Firms to Analyze Social Media After Jan. 6 Failure*, Wall St. J. (Aug. 15, 2021), https://www.wsj.com/articles/homeland-security-considers-outside-firms-to-analyze-social-media-after-jan-6-failure-11629025200?mod=rss_Technology.

[176] *See* https://www.dhs.gov/CP3; https://www.dhs.gov/science-and-technology/news/2020/02/24/news-release-dhs-selects-university-nebraska-omaha-lead-terrorism-research; https://www.unomaha.edu/ncite/. Although it involved network analysis and not AI, prior DHS grants funded development of datasets of terrorist groups that can predict which organizations are likely to increase in lethality. *See* https://www.start.umd.edu/about-baad.

the reporting of tips regarding terrorist activity.[177] In addition, DARPA's Memex program, which involved online search technology linking terrorists and human trafficking operations, was then used by MIT researchers to develop an AI-based tool.[178]

Such recognition is certainly not limited to the United States. The United Nations issued an in-depth report in 2021 about the use of AI to combat TVEC on social media, describing limits and human rights concerns for such use and identifying applications besides automated detection and takedown, including: (1) predictive analytics for terrorist activity; (2) identifying red flags of radicalization; (3) countering terrorist and violent extremist narratives; and (4) managing heavy data analysis demands.[179] The report provides many examples of public and private efforts in each area, such as the European Union's funding of a project, RED-Alert, to develop new content monitoring and analysis tools,[180] and the United Kingdom's work to develop technology to identify ISIS propaganda videos.[181] A report by the Global Network on Extremism and Technology (GNET), an academic research initiative, also provides examples of how governments in several countries are using AI tools for addressing TVEC online.[182]

As with CSAM, collaborative efforts in this space are significant. The Global Internet Forum to Counter Terrorism (GIFCT) is a non-governmental entity designed to prevent terrorists and violent extremists from exploiting digital platforms. Founded by several large tech firms in 2017, GIFCT created a shared industry database of hashes of terrorist propaganda to support coordinated takedown of such content.[183] GIFCT has expanded its membership, became an independent, non-profit organization, and is now working to broaden its database in line with human rights and privacy considerations.[184] It has also issued reports on, among other things,

---

[177] *See* https://www.unomaha.edu/ncite/news/2021/10/ncite-researchers-win-prevention-grant.php. Sponsored by the State Department, the RAND Corporation delved into the utility and the ethical and legal challenges posed by government use of bots to counter radicalization. *See* William Marcellino, et al., *Counter-Radicalization Bot Research*, RAND Corp. (2020), https://www.rand.org/pubs/research_reports/RR2705.html.

[178] *See* Kylie Foy, *Artificial intelligence shines light on the dark web*, MIT News (May 13, 2019), https://news.mit.edu/2019/lincoln-laboratory-artificial-intelligence-helping-investigators-fight-dark-web-crime-0513.

[179] United Nations Office of Counter-Terrorism, *Countering Terrorism Online with Artificial Intelligence* (2021), https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf. *See also* Kathleen McKendrick, *Artificial Intelligence Prediction and Counterterrorism*, Chatham House (2019), https://www.chathamhouse.org/sites/default/files/2019-08-07-AICounterterrorism.pdf.

[180] *See* https://cordis.europa.eu/project/id/740688.

[181] *See* https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online; https://faculty.ai/ourwork/identifying-online-daesh-propaganda-with-ai/.

[182] *See* Marie Schroeter, *Artificial Intelligence and Countering Violent Extremism: A Primer*, Global Network on Extremism and Technology (Sep. 2020), https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/.

[183] *See* https://gifct.org/tech-innovation/.

[184] *See id.*; GIFCT, *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps* (Jul. 2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-TaxonomyReport-2021.pdf. This

positive online interventions and a gap analysis looking at technical requirements for smaller platforms.[185] Also working closely with GIFCT is Tech Against Terrorism (TAT), a UN-sponsored initiative promoting information-sharing between governments and the tech sector.[186]

Besides using the GIFCT database, most major platforms deploy other automated methods to address TVEC. Facebook reportedly uses AI, combined with manual review, to attempt to understand text that might be advocating for terrorism, find and remove terrorist "clusters," and detect new accounts from repeat offenders.[187] YouTube and TikTok report using machine learning or other automated means to flag extremist videos, and Twitter indicates that it uses machine learning and human review to detect and suspend accounts responsible for TVEC.[188] Moonshot (a tech company) and Google's Jigsaw use the "Redirect Method," which uses AI to identify at-risk audiences and provide them with positive, de-radicalizing content, including pursuant to Google searches for extremist content.[189]

The efficacy and effects of the platforms' AI tools are — once again — dubious or unknown given relative lack of transparency and access to data,[190] and their potential for exacerbating bias

---

expansion effort is intended to deal with the under-representation of far-right extremists in the database, which has been the subject of critique. *See, e.g.*, Bharath Ganesh, *How to Counter White Supremacist Extremists Online*, Foreign Policy (Jan. 28, 2021), https://foreignpolicy.com/2021/01/28/how-to-counter-white-supremacist-extremists-online/.

[185] *See* GIFCT, *Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 2* (Jul. 2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI2-2021.pdf; Tech Against Terrorism and GIFCT, *Technical Approaches Working Group* (Jul. 2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf. *See also* Erin Saltman, et al., *New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis*, Studies in Conflict & Terrorism (2021), https://doi.org/10.1080/1057610X.2021.1888404.

[186] *See* https://www.techagainstterrorism.org/.

[187] *See* Erin Saltman, *Countering terrorism and violent extremism at Facebook: Technology, expertise and partnerships*, Observer Research Foundation (Aug. 27, 2020), https://www.orfonline.org/expert-speak/countering-terrorism-and-violent-extremism-at-facebook/. The importance of mapping networks of extremists across platforms in order to disrupt their reach has been studied by Google's Jigsaw and others. *See* Beth Goldberg, *Hate "Clusters" Spread Disinformation Across Social Media. Mapping Their Networks Could Disrupt Their Reach*, Jigsaw (Jul. 28, 2021), https://medium.com/jigsaw/hate-clusters-spread-disinformation-across-social-media-995196515ca5.

[188] *See* https://blog.youtube/news-and-events/more-information-faster-removals-more/; https://www.tiktok.com/safety/resources/tiktok-transparency-report-2021-q-2?lang=en; https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center.

[189] *See, e.g.*, Moonshot CVE, *Social Grievances and Violent Extremism in Indonesia* (2020), https://moonshotteam.com/resource/indonesia-social-grievances-and-violent-extremism/; https://jigsaw.google.com/issues/. *See also* Schroeter, *supra* note 182 (discussing how search engines can adjust algorithms to direct people away from extremist content).

[190] The OECD has issued reports on TVEC-related platform transparency, finding some recent improvement. OECD, *Transparency Reporting on Terrorist and Violent Content Online* (Jul. 2021), https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-8af4ab29-en.htm. GIFCT, too, has been criticized for lack of transparency. *See, e.g.*, Chloe Hadavas, *The Future of Free Speech Online May Depend on This Database*, Slate (Aug. 13, 2020), https://slate.com/technology/2020/08/gifct-

is discussed below in Section IV. As is discussed in that section, a key source of bias is the disparate or unknown performance of natural language processing on languages other than formal English, which may be analyzed as part of these efforts. While these tools, paired with human oversight, do catch some TVEC, in at least some cases these traps are more like sieves. For example, despite Facebook's admitted role in the Myanmar military's genocidal campaign in 2018 against a minority group, and despite corrective steps, its algorithms continued to amplify the military's post-coup propaganda, including incitement to violence; hateful content such as threats of murder and rape have continued into late 2021.[191]

Social media platforms and search engines are not the only places online to find TVEC. Violent extremists also find havens in messaging apps and gaming platforms, which in turn use automated tools for detection.[192] To further evade detection, extremists have also used other online sources of communication, including conference dial-in services, hospitality platforms for room bookings, and transportation applications.[193] Presumably, such services do not have the same capacity as large social media platforms and search engines to detect the presence of extremists, even assuming we would want them to collect detailed information on their users.

Academic researchers have also been studying detection methods for TVEC on social media and elsewhere. A recent literature review found a need for publicly available and unbiased datasets, a need for validation techniques to evaluate the datasets, a current research tendency to focus on ISIS ideology, and that deep learning-based methods outperformed other techniques.[194] Another

---

content-moderation-free-speech-online.html; Brittan Heller, *Combating Terrorist-Related Content Through AI and Information Sharing*, Transatlantic Working Group (Apr. 26, 2019), https://www.ivir nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf.

[191] *See* Global Witness, *Facebook approves adverts containing hate speech inciting violence and genocide against the Rohingya* (Mar. 20, 2022), https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech; Sam Neil and Victoria Milko, *Hate speech in Myanmar continues to thrive on Facebook*, AP News (Nov. 18, 2021), https://apnews.com/article/technology-business-middle-east-religion-europe-a38da3ccd40ffae7e4caa450c374f796; Global Witness, *Algorithm of harm: Facebook amplified Myanmar military propaganda following coup* (Jun. 23, 2021), https://www.globalwitness.org/en/campaigns/digital-threats/algorithm-harm-facebook-amplified-myanmar-military-propaganda-following-coup/; Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, The New York Times (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html. The problem is not limited to a single country. *See, e.g.*, Jasper Jackson, et al., *Facebook accused by survivors of letting activists incite ethnic massacres with hate and misinformation in Ethiopia*, The Bureau of Investigative Journalism (Feb. 20, 2022), https://www.thebureauinvestigates.com/stories/2022-02-20/facebook-accused-of-letting-activists-incite-ethnic-massacres-with-hate-and-misinformation-by-survivors-in-ethiopia; Mubashar Hasan, et al., *How Facebook Fuels Religious Violence*, Foreign Policy (Feb. 4, 2022), https://foreignpolicy.com/2022/02/04/facebook-tech-moderation-violence-bangladesh-religion/?tpcc=recirc_latest062921.

[192] *See* Carl Miller and Shiroma Silva, *Extremists using video-game chats to spread hate*, BBC News (Sep. 23, 2021), https://www.bbc.com/news/technology-58600181.

[193] *See* Erin Saltman, *Challenges in Combating Terrorism and Extremism Online*, Lawfare (Jul. 11, 2021), https://www.lawfareblog.com/challenges-combating-terrorism-and-extremism-online.

[194] *See* Mayur Gaikwad, et al., *Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools*, IEEE Access, vol. 9, pp. 48364-48404 (2021)

recent study noted similar concerns and added the lack of a commonly accepted definition of TVEC, the constant evolution of extremist behavior, and the need for ethical guidelines.[195]

Considering that the same extremist group may use multiple types of platforms to recruit and radicalize, that terrorist methods change, and that definitions and datasets are problematic, what seems clear is that automated tools have a long way to go in this area. Per the broader discussion below, they must be coupled with appropriate collaboration, human oversight, and a nuanced understanding of contextual and cultural difference, all while somehow striking the right balance of free speech, privacy, and safety.[196]

F. Disinformation campaigns coordinated by inauthentic accounts or individuals to influence United States elections

The Technology Engagement Team (TET) of the State Department's Global Engagement Center (GEC) defends against foreign disinformation and propaganda by leading efforts to address the problem via technological innovation. In cooperation with foreign partners, private industry, and academia, its goal is to identify, assess, and test such technologies, which often involve AI and efforts to address election-related disinformation.[197] Further, the Cybersecurity and Infrastructure Security Agency of DHS is responsible for the security of domestic elections and engages in substantial work against election-related disinformation. The Commission suggests that these agencies are best positioned to advise Congress on federal agency efforts in this area.

Several substantial reports have addressed inadequate platform efforts to address election-related disinformation, including the limited assistance of AI tools. In 2021, the Election Integrity Partnership published a lengthy report on misinformation and the 2020 election, concluding, among other things, that platform attempts to use AI to label content were flawed because the AI tools could not "distinguish false or misleading content from general election-related

---

(noting bias in terms of which ideologies, events, or organizations are included in datasets), https://doi.org/10.1109/ACCESS.2021.3068313. *See also* Sara M. Abdulla, *Terrorism, AI, and Social Media Research Clusters*, Center for Security and Emerging Technology (Nov. 2021), https://cset.georgetown.edu/publication/terrorism-ai-and-social-media-research-clusters/.

[195] Miriam Fernandez and Harith Alani, *Artificial Intelligence and Online Extremism: Challenges and Opportunities*, in Predictive Policing and Artificial Intelligence 131-62 (John McDaniel and Ken Pease, eds.) (2021) (also noting biases involving geographical location, language, and terminology), https://oro.open.ac.uk/69799/1/Fernandez_Alani_final_pdf.pdf. The definitional problem and other issues were raised in a 2020 joint letter from human rights groups to GIFCT. *See* https://www hrw.org/news/2020/07/30/joint-letter-new-executive-director-global-internet-forum-counter-terrorism#.

[196] *See, e.g.*, United Nations Office of Counter-Terrorism, *supra* note 179; Saltman, Lawfare, *supra* note 193; Jonathan Schnader, *The Implementation of Artificial Intelligence in Hard and Soft Counterterrorism Efforts on Social Media*, Santa Clara High Tech. L. J. 36:1 (Feb. 2, 2020), https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1647&context=chtlj.

[197] *See* https://www.state.gov/bureaus-offices/under-secretary-for-public-diplomacy-and-public-affairs/global-engagement-center/technology-engagement-team; https://www.state.gov/programs-technology-engagement-team/.

commentary."[198] Further, a recent ProPublica and Washington Post investigation — for which researchers relied in part on machine learning techniques – found that Facebook played a critical role in spreading false narratives about the election immediately before the January 6, 2021, siege of the United States Capitol.[199] Park Advisors, a State Department contractor working with GEC, issued a 2019 report that discussed the mixed results from platform attempts — including via the use of AI — to counter this problem in connection with recent elections.[200]

For several years, academic researchers such as University of Southern California Professor Emilio Ferrara have been using AI, sometimes with government funding, to study election-related disinformation, despite limited data available from platforms other than Twitter. In one recent study, focused on Twitter and the 2020 Presidential election, the results implied that platform efforts to limit malicious groups were not effective against those groups' evasive actions, such that "rethinking effective platform interventions is needed."[201] Another recent study involving Twitter and the 2020 election found that bots were still responsible for significant manipulation but that, as compared to the 2016 election, a shift had occurred from foreign to domestic sources.[202] Other recent studies propose platform-agnostic techniques to detect coordinated accounts or operations based on social media content or behavior.[203] Another

---

[198] Center for an Informed Public, Digital Forensic Research Lab, Graphika, & Stanford Internet Observatory, *The Long Fuse: Misinformation and the 2020 Election*, Stanford Digital Repository: Election Integrity Partnership v1.2.0 at 212 (2021), https://purl.stanford.edu/tr171zs0069. Further, to the extent that election-related disinformation often involves bots or deepfakes, the same detection problems exist in this context as they do for bots and deepfakes generally.

[199] *See* Craig Silverman, et al., *Facebook groups topped 10,000 daily attacks on election before Jan. 6, analysis shows*, The Washington Post (Jan. 4, 2022), https://www.washingtonpost.com/technology/2022/01/04/facebook-election-misinformation-capitol-riot/; Jeremy B. Merrill, *How ProPublica and The Post researched posts of Facebook groups*, The Washington Post (Jan. 4, 2022), https://www.washingtonpost.com/technology/2022/01/04/facebook-propublica-post-jan6-methodology/. *See also* Tech Transparency Project, *A Year After Capitol Riot, Facebook Remains an Extremist Breeding Ground* (Jan. 4, 2022), https://www.techtransparencyproject.org/articles/year-after-capitol-riot-facebook-remains-extremist-breeding-ground.

[200] *See* Nemr and Gangware, *supra* note 79.

[201] Karishma Sharma, et al., *Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 U.S. Presidential Election* (Oct. 20, 2021), https://arxiv.org/pdf/2107.08319.pdf.

[202] *See* Ho-Chun Herbert Chang, et al., *Social Bots and Social Media Manipulation in 2020: The Year in Review*, (Feb. 16, 2021), https://arxiv.org/pdf/2102.08436.pdf. *See also* William Marcellino, et al., *Human–machine detection of online-based malign information*, RAND Corporation (2020), https://www.rand.org/pubs/research_reports/RRA519-1 html.

[203] Karishma Sharma, et al., *Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours* (Aug. 2021), https://dl.acm.org/doi/pdf/10.1145/3447548.3467391; Steven T. Smith, et al., *Automatic detection of influential actors in disinformation networks*, PNAS 118 (4) (Jan. 26, 2021), https://www.pnas.org/content/118/4/e2011216118; Meysam Alizadeh, et al., *Content-based features predict social media influence operations*, Sci. Adv. 6: eabb5824 (Jul. 2020), https://www.science.org/doi/10.1126/sciadv.abb5824.

study showed that one can detect disinformation websites by looking not at perceptible content but at a website's infrastructure features.[204]

Besides trying to detect particular individuals and accounts that distribute election-related disinformation, AI can also be harnessed for related goals. For example, it can be used to map out communities responsible for such harm. The social media monitoring company Graphika engages in such efforts,[205] issuing multiple reports on foreign and domestic actors engaged in election-related disinformation campaigns across many platforms.[206] Looking beyond social media and big technology companies, the Wikimedia Foundation acted to support editors and community oversight of Wikipedia by investing in AI tools to counter election-related disinformation.[207] These tools included techniques to categorize and measure new content, identify unverified statements, and detect fake accounts.[208]

G. Sale of counterfeit products

In January 2020, DHS issued a report finding that private sector efforts, including those of e-commerce platforms, "have not been sufficient to prevent the importation and sale of a wide variety and large volume of counterfeit and pirated goods to the American public."[209] The report describes the efforts of the National Intellectual Property Rights Coordination Center (IPR Center) to form the Anti-Counterfeiting Consortium to Identify Online Nefarious Actors (ACTION), which intends to increase "[s]haring of risk automation techniques allowing ACTION members to create and improve on proactive targeting systems that automatically monitor online platform sellers for counterfeits and pirated goods."[210] Information collected later by the IPR Center indicated that some platforms use automated systems to verify third-party seller information and identify prohibited items.[211] Although the efficacy of these systems is unknown, platforms report undertaking some of the following efforts:

---

[204] *See* Austin Hounsel, et al., *Identifying Disinformation Websites Using Infrastructure Features*, USENIX (Sep. 11, 2020), https://www.usenix.org/conference/foci20/presentation/hounsel.

[205] *See* Jean-Baptiste Jeangène Vilmer, *Information Defense* at 24, The Atlantic Council (Jul. 2021), https://www.atlanticcouncil.org/wp-content/uploads/2021/07/Information-Defense-07.2021.pdf.

[206] *See, e.g.*, Graphika, *Posing as Patriots* (Jun. 2021), https://public-assets.graphika.com/reports/graphika_report_posing_as_patriots.pdf; Graphika, *Ants in a Web* (May 2021), https://public-assets.graphika.com/reports/graphika_report_ants_in_a_web.pdf.

[207] *See* https://wikimediafoundation.org/news/2020/10/30/how-wikipedia-is-preparing-for-election/.

[208] *Id.*

[209] Department of Homeland Security, *Combating Trafficking in Counterfeit and Pirated Goods* at 5 (Jan. 24, 2020), https://www.dhs.gov/sites/default/files/publications/20_0124_plcy_counterfeit-pirated-goods-report_01.pdf.

[210] *Id.* at 31.

[211] *See* Morgan Stevens, *National IPR Center Report Highlights Industry Adoption of Anti-Counterfeit Measures*, Center for Data Innovation (Oct. 13, 2021), https://datainnovation.org/2021/10/national-ipr-center-report-highlights-industry-adoption-of-anti-counterfeit-measures/. The IPR Center report itself is not publicly available.

- eBay has indicated it uses automated filters, including filters based on keywords, image recognition and machine learning, to flag or block problematic items, as well as to review seller information.[212]

- Etsy has indicated it started increasing its investments into automated tools, including machine learning, to detect counterfeits and other "handmade violations."[213]

- Facebook has indicated it uses automated systems, some based on machine learning, to review ads, Marketplace listings, and other content to block possible counterfeits, looking at "signals such as brand names, logos, keywords, prices, [and] discounts."[214]

- Alibaba has indicated it uses artificial intelligence in its anti-counterfeiting efforts and also started the Alibaba Anti-Counterfeiting Alliance, which includes hundreds of brands.[215]

It is unclear whether and to what extent any other social media platforms — like TikTok — are using AI or other tools to limit facilitation of off-platform sales of counterfeit goods.[216]

At least one research team has proposed an innovative system to catch counterfeits online using a clustering algorithm, among other things.[217] We could not find other academic research on this subject, suggesting that this may be an area for greater focus. Finally, it is also worth noting that some companies have developed AI tools to detect counterfeit items in the physical world.[218]

IV. RECOMMENDATIONS

The development and deployment of automated tools to address online harms will continue with or without federal encouragement. But misuse or over-reliance on these tools can lead to poor results that can serve to cause more harm than they mitigate. For this reason, Congress, government agencies, platforms, scientists, and others should focus on appropriate safeguards.

---

[212] *See* https://www.ebaymainstreet.com/issues/ebay-community-protection.

[213] *See* Corrine Pavlovic, *Our Commitment to the Trust and Safety of the Etsy Marketplace*, Etsy News Blog (Apr. 29, 2021), https://blog.etsy.com/news/2021/our-commitment-to-the-trust-and-safety-of-the-etsy-marketplace/.

[214] *See* https://www.facebook.com/business/tools/anti-counterfeiting/guide.

[215] *See* Adam Najberg, *Alibaba, Partners Notched Strong IPR Protection Gains in 2020*, Alizila (Mar. 26, 2021), https://www.alizila.com/alibaba-partners-notched-strong-ipr-protection-gains-in-2020/.

[216] *See, e.g.*, Megan Graham, *TikTok teens are obsessed with fake luxury products*, CNBC News (Mar. 1, 2020), https://www.cnbc.com/2020/02/29/tiktok-teens-are-obsessed-with-fake-luxury-products.html.

[217] *See* Patrick Arnold, et al., *Semi-automatic identification of counterfeit offers in online shopping platforms*, Journal of Internet Commerce 15(1): 59-75 (Jan. 2. 2016), https://dbs.uni-leipzig.de/file/product-counterfeits-15332861.2015.pdf.

[218] *See, e.g.*, Entrupy, *State of the Fake: 2020 Edition* (2020), https://www.mannpublications.com/fashionmannuscript/2020/09/11/entrupy-state-of-the-fake-2020-edition/; Donna Dillenberger, *Pairing AI with Optical Scanning for Real-World Product Authentication*, IBM Research Blog (May 23, 2018), https://www.ibm.com/blogs/research/2018/05/ai-authentication-verifier/.

That difficult task requires answering a host of questions for any given harm or innovation, such as who built the tool, how, and why. Others involve how the harm is being defined and who is using the tool in what environment and for what reason. Still others involve how well the tool actually works, its real-world impacts, who has authority to get answers to these questions, and who is accountable for unfair, biased, or discriminatory outcomes.

With the intense focus on the role and responsibility of social media platforms, it is often lost that other private actors — as well as government agencies — could use AI to address these harms. Many parts of the online ecosystem provide conduits for illegal or toxic content.[219] These actors include not just search engines, gaming platforms, messaging apps, marketplaces and app stores,[220] but also those at other layers of the tech stack such as internet service providers, content distribution networks, domain registrars, cloud providers, and web browsers. Via automated tools or otherwise, these companies exercise remarkable control, able to block or slow access to websites and other services, change what information consumers see, and warn people or redirect them from certain content.[221] The benefits and risks of having such actors address harmful content are beyond this report's scope, but they demand attention when approaching legal or technical solutions in this area.[222] This attention involves not merely a law's coverage or technological feasibility but also the extent to which we are comfortable with certain public or private actors wielding these powerful tools.[223]

As for the platforms, extensive accounts and in-depth analyses exist regarding their use of automated tools to address harmful content, as well as the problems with and limitations of such

---

[219] *See* Jenna Ruddock and Justin Sherman, *Widening the Lens on Content Moderation*, Joint PIJIP/TLS Research Paper Series 69 (Jul. 2021) (mapping the "online information ecosystem" beyond the "last mile" of social media), https://digitalcommons.wcl.american.edu/research/69.

[220] Yet another example is podcasting. One researcher is using AI to study misinformation, including election-related content, in podcasts, noting that it would be expensive and difficult to use such tools at scale, especially given the way podcasts are distributed. *See* Valerie Wirtschafter, *The challenge of detecting misinformation in podcasting*, Brookings Techstream (Aug. 25, 2021), https://www.brookings.edu/techstream/the-challenge-of-detecting-misinformation-in-podcasting/. *See also* Valerie Wirtschafter and Chris Meserole, *Prominent political podcasters played big role in spreading the 'Big Lie,'* Brookings Techstream (Jan. 4, 2022), https://www.brookings.edu/techstream/prominent-political-podcasters-played-key-role-in-spreading-the-big-lie/.

[221] *See* Ruddock and Sherman, *supra* note 219.

[222] *See* Corrine Cath and Jenna Ruddock, *One Year After the Storming of the US Capitol, What Have We Learned About Content Moderation Through Internet Infrastructure?*, Tech Policy Press (Jan. 6, 2022), https://techpolicy.press/one-year-after-the-storming-of-the-us-capitol-what-have-we-learned-about-content-moderation-through-internet-infrastructure/?s=03; Karl Bode, *Winding Down Our Latest Greenhouse Panel: Content Moderation At The Infrastructure Layer*, Tech Policy Greenhouse (Oct. 8, 2021), https://www.techdirt.com/articles/20211005/06472747699/winding-down-our-latest-greenhouse-panel-content-moderation-infrastructure-layer.shtml; Joan Donovan, *Navigating the Tech Stack: When, Where and How Should We Moderate Content?*, Centre for Int'l Gov. Innovation (2019), https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content/; Annemarie Bridy, *Remediating Social Media: A Layer-Conscious Approach*, 24 B.U. J. Sci. & Tech. L. 193 (2018), https://www.bu.edu/jostl/files/2018/10/Bridy-%E2%80%94-FINAL.pdf.

[223] "It's not 'What will AI do to us on its own?' It's 'What will the powerful do to us with the AI?'" Zeynep Tüfekçi, *Coded Bias*, directed by Shalini Kantayya. New York: 7th Empire Media, 2020.

use.[224] Regardless of the tools at issue, it is important to recognize, as Tarleton Gillespie has explained, that this moderation of harmful and other content is "central to what platforms do, not peripheral" and "is, in many ways, *the* commodity that platforms offer."[225] The platforms each provide an organized, curated experience of online information, often using AI tools to maximize engagement.[226] To focus only on how they may use AI to clean up the resulting mess — to moderate content they allowed users to post — can obscure the commercial reasons why and how that content got there in the first place.[227] Professor Sarah T. Roberts of the University of California, Los Angeles, who coined the phrase "commercial content moderation," called its role "fundamentally a matter of brand protection."[228] Thus, in the words of Professor Olivier Sylvain, the use of AI for content moderation "is more likely an incident of these companies' overt industrial designs on the control and consolidation of the distribution of user information."[229]

No matter who is responsible for these harms, though, the question that Congress has asked us to address is whether AI can help ameliorate them. It seeks recommendations on reasonable policies, practices, and procedures for this use of AI and for any legislation that may advance it. The following sections of this report attempt to provide such recommendations, starting with a discussion of why advancing AI for these purposes is not always the most constructive thing to do.

---

[224] *See, e.g.*, Coalition to Fight Digital Deception ("CFDD"), *Trained for Deception: How Artificial Intelligence Fuels Online Disinformation* (Sep. 2021), https://www.fightdigitaldeception.com/; Carey Shenkman, et al., *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*, Center for Democracy and Technology (May 2021), https://cdt.org/insights/do-you-*See*-what-i-*See*-capabilities-and-limits-of-automated-multimedia-content-analysis/; Tarleton Gillespie, *Content moderation, AI, and the question of scale*, Big Data & Society (Aug. 21, 2020), https://doi.org/10.1177/2053951720943234; Robert Gorwa, et al., *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, Big Data & Society (Feb. 28, 2020), https://doi.org/10.1177/2053951719897945; Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, New America Open Technology Institute (Jul. 15, 2019), https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/.

[225] Tarleton Gillespie, Custodians of the Internet at 13 (2018).

[226] *See, e.g.*, Karen Hao, *How Facebook got addicted to spreading disinformation*, MIT Tech. Rev. (Mar. 11, 2021), https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[227] *See* Shoshana Zuboff, *The Coup We Are Not Talking About*, The New York Times (Jan. 29, 2021) (referring to content moderation as "a last resort" and "a public-relations operation" meant to "minimize the risk of user withdrawal or to avoid political sanctions"), https://www.nytimes.com/2021/01/29/opinion/sunday/facebook-surveillance-society-technology html; Gillespie, Custodians of the Internet, *supra* note 225 at 198 (content moderation improvements "are all are just tweaks" that platforms may be pressured into making "while preserving their ability to conduct business as usual").

[228] Sarah T. Roberts, *Digital detritus: 'Error' and the logic of opacity in social media content moderation*, First Monday 23: 3-5 (Mar. 2018), http://dx.doi.org/10.5210/fm.v23i3.8283.

[229] Olivier Sylvain, *Recovering Tech's Humanity*, 119 Colum. L. Rev. F. 252, 265 (2019), https://ir.lawnet.fordham.edu/faculty_scholarship/1088. *See also* Joan Donovan, *Trolling for Truth on Social Media*, Scientific American (Oct. 12, 2020), https://www.scientificamerican.com/article/trolling-for-truth-on-social-media/.

A. Avoiding over-reliance

AI detection tools for the harms discussed here are blunt instruments.[230] For several reasons, their use can result in false positives and false negatives. One can adjust variables to catch more or less of a given type of content, but trade-offs are inevitable. For example, blocking more content that might incite extremist violence (e.g., via detection of certain terms or imagery) can result in also blocking members of victimized communities from discussing how to address such violence. This fact explains in part why each specified harm needs individual consideration; the trade-offs we may be willing to accept may differ for each one.[231] But what the public is willing to accept may not matter if only those developing and deploying these tools get to decide what types and levels of failure are tolerable, whether and how to assess risks and impacts, and what information is disclosed.

Built-in imprecision

Many of the AI systems built to detect particular kinds of content are "trained" to work by researchers who have fed it a set of examples that they have classified in various ways.[232] These datasets and classifications allow the system to predict whether a new example fits a given classification. For example, researchers might use a database of animal images in which some are labeled as "cats" and others as "not cats." Then the researchers may feed in new images and ask the system to decide which ones are "cats." For the system to work well, the dataset must be sufficiently big, accurate, and representative, so that no types of cats are excluded and no other animals are misbranded as feline. But the AI doesn't actually understand what a "cat" is. It's just trying to do some math. So, if the cats in the dataset include only cats with pointy ears, the system may not identify ones whose ears fold down. And if the system is trained to identify "cats" only by pointy ears and whiskers, then rabbits and foxes may be shocked to learn that they

[230] Despite marketing pitches that trumpet the use of AI, some of these tools may not be AI at all and may not even be all that automated, relying instead on something as simple as spreadsheets or on the insertion of an interface that masks underlying human labor.

[231] *See, e.g.*, United Nations, *supra* note 127 at 43; Nafia Chowdhury, *Automated Content Moderation: A Primer*, Stanford Cyber Policy Center (Mar. 19, 2022), https://cyber fsi.stanford.edu/news/automated-content-moderation-primer; Samidh Chakrabarti, Twitter Post (Oct. 3, 2021) ("This is where the rubber hits the road. What is the acceptable tradeoff between benign and harmful posts? To prevent X harmful posts from going viral, would you be willing to prevent Y benign posts from going viral? No easy answers."), https://twitter.com/samidh/status/1444544160518733824.

[232] This work is not all done by scientists. Some big technology companies use low-paid microworkers, sometimes refugees in other parts of the world, to help with the huge amount of data training needed for these systems to work. *See* Karen Hao and Andrea Paola Hernández, *How the AI industry profits from catastrophe*, MIT Tech. Rev. (Apr. 20, 2022), https://www.technologyreview.com/2022/04/20/1050392; Julian Posada, *Family Units*, Logic (Dec. 25, 2021), https://logicmag.io/beacons/family-units/; Phil Jones, *Refugees help power machine learning advances at Microsoft, Facebook, and Amazon*, Rest of World (Sep. 22, 2021), https://restofworld.org/2021/refugees-machine-learning-big-tech/.

are "cats," too. A poorly built AI system for identifying cat imagery might thus do much worse at this task than a human toddler, but it can do it a whole lot faster.[233]

For an AI tool to recognize particular online content as harmful, the calculus is much more complex than a binary question about an animal. The availability of robust, representative, and accurate datasets is a serious problem in developing these tools, as noted above with respect to harassment and TVEC. Another problem — one more inherent to machine learning — is that these tools are trained on previously identified data and thus are generally bad at detecting new phenomena.[234] Platforms cannot solve this problem merely by adding data over time, because "more data is not the same as more varied data" and because no dataset can ever include all new examples.[235] Many errors with these tools will also occur because of their probabilistic nature.[236] Beyond technological limitations, the operation of these tools is also subject to platform moderation policies that dictate what happens to particular content but that may be flawed in substantial ways.

The theoretical cat detector described above also reflects the fact that an AI tool is measuring data that serves merely as a proxy for what it is really trying to identify.[237] One reason that social media platforms have often failed to detect certain types of harmful content, like harassment, is that their automated tools are built to ignore meaning and context, focusing instead on measurable patterns of data that are based on past content moderation decisions and practices.[238] Such proxies are thus given power to stand in for something real and complex in the world.[239]

---

[233] After FTC staff imagined this system, Google Research introduced StylEx, an approach for visual explanation of classifiers. It allows someone to disentangle attributes and see what leads the model to make its decisions. It demonstrates this ability by showing how it distinguishes cats and dogs; one attribute making it less likely the model will choose "cat" is folded-down ears. *See* https://ai.googleblog.com/2022/01/introducing-stylex-new-approach-for html.

[234] *See* Gillespie, Custodians of the Internet, *supra* note 225 at 105-110; Nicolas P. Suzor, Lawless: The Secret Rules That Govern Our Digital Lives at 155 (2019). *See also* Cade Metz, The Genius Makers at 268-69 (2021) (describing the failure of Facebook's automated systems to flag the livestreaming of the deadly Christchurch incident "because it didn't look like anything those systems had been trained to recognize"); Neal Mohan, *Inside Responsibility: What's next on our misinfo efforts*, YouTube Blog (Feb. 17, 2022) (discussing YouTube challenges), https://blog.youtube/inside-youtube/inside-responsibility-whats-next-on-our-misinfo-efforts/.

[235] Gillespie, Custodians of the Internet, *supra* note 225 at 107; Cathy O'Neil, Weapons of Math Destruction at 204 (2016) ("Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide.").

[236] *See* evelyn douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality & Probability*, 121 Colum. L. Rev. 759 (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679607.

[237] *See* Gillespie, Custodians of the Internet, *supra* note 225 at 105-110.

[238] *Id.* at 104.

[239] *See* Dylan Mulvin, Proxies: The Cultural Work of Standing In at 13, 78, 106 (2021). *See also* Anya E.R. Prince and Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 Iowa L. Rev. 1257 (2020), https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data.

Context and meaning

That designing AI tools involves the removal of context likely explains, at least in part, why these tools often have yet another serious problem: they aren't good at understanding context, meaning, and intent, which can be key to deciding whether a piece of content is unlawful, against platform policy, or otherwise harmful.[240] An oft-used illustration is the phrase "I'm going to kill you," which could be either a violent threat or a jocular reply to a friend. Automated detection tools are especially poor judges of context for content that has fluid definitions[241] or where meanings may shift depending on regional, cultural, and linguistic differences. As the Surgeon General and others have argued, platforms need to "increase staffing of multilingual content moderation teams and improve the effectiveness of machine learning algorithms in languages other than English since non-English-language misinformation continues to proliferate."[242]

Bias and discrimination

The problems with automated detection tools described above, including unrepresentative datasets, faulty classifications, failure to identify new phenomena, missing context, and flawed design, can lead to biased,[243] discriminatory, or unfair outcomes. The tools can thus exacerbate some of the very harms they are intended to address and hurt some of the very people they are supposed to help.[244] This well-recognized fact is why it is so important that the use of these tools be more transparent, open to research, and subject to mechanisms for accountability.

---

[240] *See, e.g.*, Slaughter, *supra* note 13 at 13 (discussing facial recognition); Shenkman, *supra* note 224; Hannah Bloch-Wehba, *Automation in Moderation*, 53 Cornell Int'l L. J. 41 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619; Niva Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence* at 5, Big Data & Society (Jul. 29, 2020), https://doi.org/10.1177/2053951720932296; Gillespie, Custodians of the Internet, *supra* note 225 at 105. CSAM is a counterexample as to which context and intent are irrelevant.

[241] CFDD, *supra* note 224 at 10-13.

[242] Vivek H. Murphy, *Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment* at 12 (2021), https://www hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf; *See* United Nations, *supra* note 127 at 44-46 (also noting the problem of detecting sarcasm and irony); Singh, *supra* note 224 at 34.

[243] In this context, "bias" is often used as an umbrella term referring to unfairness or injustice infecting automated systems. Some have argued against focusing too much on technological causes, arguing that all data is biased and that power imbalances shape the data being used in these systems. *See* Milagros Miceli, et al., *Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?*, Proc. ACM Hum.-Comput. Interact. 6, GROUP, Art. 34 (Jan. 2022), https://doi.org/10.1145/3492853.

[244] It can also hurt the bottom line. A recent survey revealed that tech companies have reported losing revenue and customers because of bias (including bias based on gender, age, and race) in AI models they employed, even though some of them tested for bias in advance. *See* Veronica Combs, *Guardrail failure: Companies are losing revenue and customers due to AI bias*, TechRepublic (Jan. 11, 2022), https://www.techrepublic.com/article/guardrail-failure-companies-are-losing-revenue-and-customers-due-to-ai-bias/.

Reflecting extensive scholarship in this area,[245] several government agencies and officials have recognized generally that AI systems can be infected by bias, have discriminatory impacts, and harm marginalized communities. In October 2021, officials from the White House Office of Science and Technology Policy (WHOSTP) called for an AI bill of rights, stating that "[t]raining machines based on earlier examples can embed past prejudice and enable present-day discrimination."[246] FTC Commissioner Rebecca Kelly Slaughter has described the same problems, pointing to faulty inputs and design as well as a lack of testing.[247] Assistant Attorney General Kristen Clarke, head of the Civil Rights Division, has also spoken about bias and discrimination in AI.[248] The National Institute of Standards and Technology published a report on identifying and managing bias in artificial intelligence, based on the same concerns.[249] In its accountability framework, the Government Accountability Office referred to AI systems developed from data reflecting "preexisting biases or social inequities."[250] These issues have also been acknowledged in Executive Orders and other documents.[251]

Bias and discrimination in AI systems have also been the subject of Congressional inquiry. For example, in a 2019 hearing, Professor Meredith Whittaker testified that bias in AI systems results from "faulty training data, problems in how the system was designed or configured, or bad or biased applications in real world contexts. In all cases it signals that the environments where a given system was created and envisioned didn't recognize or reflect on the contexts within which these systems would be deployed. Or, that those creating and maintaining these systems did not

---

[245] *See, e.g.*, Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (2019); Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (2018); Rashida Richardson, *Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities*, Berkeley Tech. L. J. 36:3 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850317.

[246] Eric Lander and Alondra Nelson, *Americans Need a Bill of Rights for an AI-Powered World*, WIRED (Oct. 8, 2021), https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/?s=03.

[247] Slaughter, *supra* note 13 at 7-14.

[248] *See* https://www.justice.gov/opa/speech/assistant-attorney-general-kristen-clarke-delivers-keynote-ai-and-civil-rights-department.

[249] *See* NIST, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST Special Publication 1270 (Mar. 2022), https://doi.org/10.6028/NIST.SP.1270.

[250] Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities* at 9, GAO-21-519SP (Jun. 2021), https://www.gao.gov/assets/gao-21-519sp.pdf.

[251] In 2020, the White House issued two documents that acknowledged problems of bias, discrimination, fairness, and privacy in AI systems. *See* Exec. Order No. 13960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*, 85 Fed. Reg. 78939 (Dec. 3, 2020) (on AI systems deployed by government agencies), https://www.govinfo.gov/content/pkg/FR-2020-12-08/pdf/2020-27065.pdf; Office of Management and Budget Memorandum M-21-06, *Guidance for Regulation of Artificial Intelligence Applications* (Nov. 17, 2020) ("OMB Memo") (on AI systems deployed outside the government), https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf. *See also* Select Committee on Artificial Intelligence, National Science and Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* at 21-26 (Jun. 2019) ("NAIRD Strategic Plan"), https://www nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf; Laurie A. Harris, *Artificial Intelligence: Background, Selected Issues, and Policy Considerations* at 41-42, CRS Report No. R46795 (May 19, 2021), https://crsreports.congress.gov/product/pdf/R/R46795.

have the experience or background to understand the diverse environments and identities that would be impacted by a given system."[252]

The use of AI in the automated detection of online harms is certainly not immune to these issues.[253] For example, large language models used to moderate online content can result in biased and discriminatory results given the flaws in those models.[254] The problem of biased and unrepresentative datasets are discussed above in connection with harassment and TVEC detection, and at least three studies specifically revealed bias in several hate speech detection models.[255] Internal Facebook documents reportedly show that its hate speech detection model operated in a way that left members of minority communities and people of color more open to abuse than, say, white men.[256] Further, the Brennan Center for Justice and the Coalition for Digital Democracy have each explored these issues and cited examples of platform use of AI in

---

[252] *Artificial Intelligence: Societal and Ethical Implications*, H. Comm. on Science, Space, and Technology, 116th Cong. (2019) (testimony of Meredith Whittaker), https://science house.gov/hearings/artificial-intelligence-societal-and-ethical-implications. *See also* NIST Special Publication 1270, *supra* note 249 at 32-33; Harris, *supra* note 251 at 10, 42; Sendhil Mullainathan and Ziad Obermeyer, *On the Inequity of Predicting A While Hoping for B*, AEA Papers and Proceedings 111: 37–42 (2021), https://doi.org/10.1257/pandp.20211078; Alex V. Cipolle, *How Native Americans Are Trying to Debug A.I.'s Biases*, The New York Times (Mar. 22, 2022), https://www.nytimes.com/2022/03/22/technology/ai-data-indigenous-ivow html.

[253] *See, e.g.*, Suzor, *supra* note 234 at 155; Shenkman, *supra* note 224 at 26-28; Gorwa, *supra* note 224 at 10-11.

[254] *See* Laura Weidinger, et al., *Ethical and social risks of harm from Language Models*, DeepMind (Dec. 2021), https://storage.googleapis.com/deepmind-media/research/language-research/Ethical%20and%20social%20risks.pdf; Emily Bender, et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in ACM Conf. on Fairness, Accountability, and Transparency, at 613-15 (Mar. 2021) (noting that size doesn't guarantee diversity), https://doi.org/10.1145/3442188.3445922; Karen Hao, *The race to understand the thrilling, dangerous world of language AI*, MIT Tech. Rev. (May 20, 2021), https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/. In late 2021, Microsoft and NVIDIA reported it had developed the largest such model trained to date but acknowledged that it is infected with bias. *See* https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/. DeepMind also announced a new language model in 2021 and, similarly, acknowledged that eliminating harmful language is an ongoing problem. *See* Will Douglas Heaven, *DeepMind says its new language model can beat others 25 times its size*, MIT Tech. Rev. (Dec. 8, 2021), https://www.technologyreview.com/2021/12/08/1041557/deepmind-language-model-beat-others-25-times-size-gpt-3-megatron/. *See also* Kate Kaye, *OpenAI's new language AI improves on GPT-3, but still lies and stereotypes*, Protocol (Jan. 27, 2022) (despite small improvements, OpenAI's model "still has tendencies to make discriminatory comments and generate false information"), https://www.protocol.com/enterprise/openai-gptinstruct.

[255] *See* Rottger, *supra* note 129; Maarten Sap, et al., *The Risk of Racial Bias in Hate Speech Detection*, in Proc. of the 57th Ann. Meeting of the Ass'n for Computational Linguistics 1668 (2019), https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf; Thomas Davidson, et al., *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, Proc. of the Third Abusive Language Workshop at the Ann. Meeting for the Ass'n for Computational Linguistics 6 (Aug. 1–2, 2019), https://arxiv.org/pdf/1905.12516.pdf. Biased outcomes are a risk generally when natural language processing is applied to languages other than formal English. *See* Patton, *supra* note 168; Su Lin Blodgett and Brendan O'Connor, *Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English* (Jun. 30, 2017), https://arxiv.org/pdf/1707.00061.pdf; Natasha Duarte, et al., *Mixed Messages? The Limits of Automated Social Media Content Analysis*, Center for Democracy & Technology (2017), https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

[256] *See* Elizabeth Dwoskin, et al., *Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show*, The Washington Post (Nov. 21, 2021), https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/?s=03.

content moderation that produced discriminatory outcomes and hurt already marginalized groups.[257]

When the use of an AI detection tool results in false positives, or overblocking, it may serve to reduce freedom of expression. This problem is especially acute when those silenced are members of historically marginalized communities. Several experts, including Stanford University Professor Daphne Keller and Emma Llansó, have written about these speech effects in connection with TVEC and other content.[258] Weighed against the risks of overblocking, of course, are the risks of underblocking, which can also implicate free expression. As Professor Citron and University of Miami Professor Mary Anne Franks have argued, online harassment acts to silence its targets, who may close social media accounts and not engage in public discourse.[259]

Evasion and attack

In the previous section, we identified several areas — bot-driven accounts, deepfakes, illegal drug sales, and violent extremism — in which bad actors find ways to avoid automated detection tools. Some of them may use their own technological tools to do so, like sophisticated techniques for media manipulation. Others may find that simpler evasion methods do the trick, such as inserting typos or using innocuous phrases or euphemisms,[260] using new slurs or special icons or

[257] *See* Brennan Center for Justice, *Double Standards in Social Media Content Moderation* (Aug. 4, 2021), https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation; CFDD, *supra* note 224 at 11-13.

[258] *See* Daphne Keller, *Making Google the Censor*, The New York Times (Jun. 12, 2017) ("no responsible technologist believes that filters can tell what speech is legal," a call even "[s]killed lawyers and judges struggle to make"), https://www nytimes.com/2017/06/12/opinion/making-google-the-censor html; Emma J. Llansó, *No amount of "AI" in content moderation will solve filtering's prior-restraint problem*, Big Data & Society 7(1) (2020), https://doi.org/10.1177/2053951720920686. *See also* Tech Against Terrorism, *GIFCT Technical Approaches Working Group Gap Analysis and Recommendations* at 32 (Jul. 2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf.

[259] Danielle Keats Citron and Mary Ann Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, U. Chi. Legal F. Vol. 2020 (2020), https://chicagounbound.uchicago.edu/uclf/vol2020/iss1/3. *See also* Sophia Smith Galer, *'This Your Girlfriend?': Videos Shaming Women for Sex Jokes Go Viral on TikTok*, Vice World News (Nov. 30, 2021), https://www.vice.com/en/article/v7ddzd/this-your-girlfriend-videos-shaming-women-for-sex-jokes-go-viral-on-tiktok; Amnesty International, *Toxic Twitter – The Silencing Effect* (March 2018), https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-5/.

[260] *See* CFDD, *supra* note 224 at 14; Ana Romero-Vicente, *Word Camouflage to Evade Content Moderation*, EU DisinfoLab (Dec. 2, 2021), https://www.disinfo.eu/publications/word-camouflage-to-evade-content-moderation/; Tommi Gröndahl, et al., *"All You Need is "Love": Evading Hate Speech Detection*, Proc. of the 11th ACM Workshop on Artificial Intelligence and Security (Nov. 5, 2018), https://arxiv.org/abs/1808.09115v3; Hao, *How Facebook Got Addicted to Spreading Disinformation*, *supra* note 226.

logos,[261] altering or covering up images,[262] adding sounds to camouflage audio tracks,[263] using ephemeral features such as Instagram Stories,[264] or switching to unmonitored channels, like some comment sections or audio chat services.[265] This constant arms race demands that those responsible for detection remain vigilant, considering and adjusting for possible and actual evasions throughout the technology's lifecycle.

Another substantial concern is that AI systems are vulnerable to hacking and manipulation.[266] DARPA, NIST, and the Department of Defense's Joint Artificial Intelligence Center are all working on projects that aim to better protect AI systems from attack.[267] But this concern — often discussed using terms like adversarial machine learning and adversarial robustness — is not limited to the military context.

It is perhaps no wonder, taking all of these factors into account, that AI is not the easy answer to addressing online harms. Few people would claim otherwise. For example, Facebook officials and employees have confirmed repeatedly that AI systems do not catch a significant percentage of harmful content. In 2018, Monika Bickert, its Head of Global Policy Management, stated that "we're a long way" from AI solving content moderation problems such as determining whether something amounts to harassment or bullying.[268] The following year, a Facebook engineer commented, "The problem is that we do not and possibly never will have a model that captures

---

[261] *See* Mark Scott, *Islamic State evolves 'emoji' tactics to peddle propaganda online*, Politico EU (Feb. 10, 2022), https://www.politico.eu/article/islamic-state-disinformation-social-media/; Sentropy Technologies, *Why is content moderation so hard?* (Oct. 1, 2020), https://medium.com/sentropy/why-is-content-moderation-so-hard-e1e16433337f.

[262] *See* The Virality Project, *Content moderation avoidance strategies* (Jul. 29, 2021), https://www.viralityproject.org/rapid-response/content-moderation-avoidance-strategies-used-to-promote-vaccine-hesitant-content.

[263] *See* Sophia Smith Galer, *Anti-Vaxxers Are Learning How To Game TikTok's Algorithm — And They're Going Viral*, Vice World News (Sep. 6, 2021), https://www.vice.com/en/article/v7ek3d/anti-vaxxers-are-learning-how-to-game-tiktoks-algorithm-and-theyre-going-viral.

[264] *Id.*

[265] *See* Elizabeth Dwoskin, et al., *Racists and Taliban supporters have flocked to Twitter's new audio service after executives ignored warnings*, The Washington Post (Dec. 10, 2021), https://www.washingtonpost.com/technology/2021/12/10/twitter-turmoil-spaces/; Sam Schechner, et al., *How Facebook Hobbled Mark Zuckerberg's Bid to Get America Vaccinated*, The Wall St. J. (Sep. 18, 2021), https://www.wsj.com/articles/facebook-mark-zuckerberg-vaccinated-11631880296.

[266] *See* Andrew J. Lohn, *Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity*, Center for Security and Emerging Technology (Dec. 2020), https://cset.georgetown.edu/publication/hacking-ai/; Ryan Calo, et al., *Is Tricking a Robot Hacking?*, U. Wash. Tech Policy Lab Legal Studies Research Paper No. 2018-05 (Mar. 28, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3150530; Pin-Yu Chen, Securing AI systems with adversarial robustness, IBM Research Blog (Dec. 15, 2021), https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness.

[267] *See, e.g.*, https://www.darpa mil/news-events/2021-12-21; https://www nccoe nist.gov/ai/adversarial-machine-learning; Will Knight, *The Pentagon Is Bolstering Its AI Systems—by Hacking Itself*, WIRED (Jul. 19, 2021), https://www.wired.com/story/pentagon-bolstering-ai-systems-hacking-itself/.

[268] *See* Alexis C. Madrigal, *Inside Facebook's Fast-Growing Content-Moderation Effort*, The Atlantic (Feb. 7, 2018), https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/.

even a majority of integrity harms, particularly in sensitive areas."[269] In 2021, an executive, Andrew Bosworth, wrote in an employee memo that moderating people's behavior in the metaverse "at any meaningful scale is practically impossible."[270]

Moreover, even as these tools become better at identifying explicitly harmful content, neither machines nor human moderators may ever be able to deal effectively with "the mass of ordinary and pervasive posts that express discriminatory sentiments in ways that threaten and silence marginalized groups."[271] Queensland University of Technology Professor Nicolas Suzor, who sits on the Facebook Oversight Board, calls such posts "[t]he internet's major abuse problem" and explains that "[u]ltimately, abuse and harassment are not just problems of content classification."[272] It's not clear that automating decisions about certain kinds of harmful content is something to which platforms or others should aspire anyway. Tarleton Gillespie argues that these decisions "are judgments of value, meaning, importance, and offense. They depend both on a human revulsion to the horrific and a human sensitivity to contested cultural values. There is, in many cases, no right answer for whether to allow or disallow, except in relation to specific individuals, communities, or nations that have debated and regulated standards of propriety and legality."[273]

B. Humans in the loop

If AI tools employed to detect harmful online content are not good or fair enough to work on their own, then an obvious and widely shared conclusion is that they need appropriate human oversight.[274] Professor Sarah T. Roberts explained that the many kinds of harmful content poorly suited for automated filters require humans "called upon to employ an array of high-level cognitive functions and cultural competencies to make decisions about their appropriateness for a site or platform."[275] Their judgment may also be constrained or distorted by the content moderation policies they are required to enforce. Given the amount of online content through which to wade, however, it is entirely implausible to put enough humans in place to monitor all

---

[269] *See* Seetharaman, *supra* note 129.

[270] *See* Adi Robertson, *Meta CTO thinks bad metaverse moderation could pose an 'existential threat,'* The Verge (Nov. 12, 2021), https://www.theverge.com/2021/11/12/22779006/meta-facebook-cto-andrew-bosworth-memo-metaverse-disney-safety-content-moderation-scale. *See also* Emily Baker-White, *Meta Wouldn't Tell Us How It Enforces Its Rules in VR, So We Ran a Test to Find Out*, BuzzFeed News (Feb. 11, 2022), https://www.buzzfeednews.com/article/emilybakerwhite/meta-facebook-horizon-vr-content-rules-test;Tanya Basu, *This group of tech firms just signed up to a safer metaverse*, MIT Tech. Rev. (Jan. 10, 2022) (describing why current AI detection tools for online harms will fare poorly in the metaverse), https://www.technologyreview.com/2022/01/20/1043843/safe-metaverse-oasis-consortium-roblox-meta/.

[271] Suzor, *supra* note 234 at 65.

[272] *Id.*

[273] Gillespie, Custodians of the Internet, *supra* note 225 at 206.

[274] *See, e.g.,*; Gillespie, Custodians of the Internet, *supra* note 225 at 107; Rachel Thomas, *Avoiding Data Disasters* (Nov. 4, 2021), https://www.fast.ai/2021/11/04/data-disasters/; CFDD, *supra* note 224 at 14; Shenkman, *supra* note 224 at 36; Singh, *supra* note 224 at 34; Google, *Removals under the Network Enforcement Law* ("Machine automation simply cannot replace human judgment and nuance."), https://perma.cc/SF24-X6ZK.

[275] Sarah T. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media (2019) at 34-35.

of it, which is why platforms generally use moderators as part of a triage system. Nonetheless, it would help if more human moderators were at work.[276] It is also true that the risk of harm from some automated tools, like those intended to catch sales of certain illegal or counterfeit goods, may be low enough — at least in terms of false positives — that a relative lack of human review is acceptable.

Simply placing moderators, trust and safety professionals, and other people in AI oversight roles is insufficient. The work is challenging and demands that moderators have adequate training, time, agency to make decisions, and workplace protections.[277] To determine exactly what makes such oversight meaningful will require more research and analysis.[278] Further, humans come with their own implicit biases, which can be exacerbated if they are poorly trained and need to make snap judgments.[279] They can also be subject to automation bias, i.e., a tendency to be overly deferential to automated decisions.[280] Teams of moderators should thus be diverse and, as already noted, understand many different cultures and languages. A report from New York University's Stern Center for Business and Human Rights provides specific recommendations for large platforms, including (1) doubling the number of moderators, (2) making them full-time employees with suitable pay and benefits, (3) expanding efforts in other countries with moderators who know local culture and language, and (4) providing good medical care and sponsoring research on health risks.[281] Some writers have also pointed out that, of course, having

---

[276] *See*, e.g.. Gillespie, Custodians of the Internet, *supra* note 225 at 198; Suzor, *supra* note 234 at 65.

[277] *See* Roberts, Behind the Screen, *supra* note 275; Andrew Strait, *Why content moderation won't save us*, in Fake AI, *supra* note 4 at 147-58 (describing platforms treating moderators as an expendable and undervalued people whose "hidden emotional labour" keeps the automated systems afloat); Ben Wagner, *Liable, but Not in Control? Ensuring Human Agency in Automated Decision-Making Systems*, Policy & Internet 11(1) (2019), https://doi.org/10.1002/poi3.198; Billy Perrigo, *Inside Facebook's African Sweatshop*, TIME (Feb. 14, 2022), https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/?s=03; Parmy Olson, *How Facebook and Amazon Rely on an Invisible Workforce*, The Washington Post (Jan. 20, 2022), https://www.washingtonpost.com/business/how-facebook-and-amazon-rely-on-an-invisible-workforce/2022/01/20/c7305bfa-79c7-11ec-9dce-7313579de434_story html.

[278] *See* Rebecca Crootof, et al., *Humans in the Loop*, Vand. L. Rev. (forthcoming 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066781; Ben Green and Amba Kak, *The False Comfort of Human Oversight as an Antidote to A.I. Harm*, Slate (Jun. 15, 2021), https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html.

[279] *See* Brennan Center for Justice, *supra* note 257 at 10-11; Green and Kak, *supra* note 278.

[280] *See* Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, (Sep. 10, 2021), http://dx.doi.org/10.2139/ssrn.3921216; Linda J. Skitka, et al., *Accountability and automation bias*, Int'l J. of Human-Computer Studies 52:4 (Apr. 2000), https://doi.org/10.1006/ijhc.1999.0349.

[281] *See* Paul M. Barrett, *Who Moderates the Social Media Giants?*, NYU Stern Center for Bus. and Human Rights (Jun. 2021), https://bhr.stern nyu.edu/tech-content-moderation-june-2020?_ga=2.195456940.1820254171.1645560371-1997396386.1645560371. *See also* Mohan, *supra* note 234 (referring to YouTube hiring moderators with understanding of regional nuances and local languages).

humans in the loop doesn't correct for harms caused by flawed AI systems; it also shouldn't serve as a way to legitimize such systems or for their operators to avoid accountability.[282]

C. Transparency and accountability

Calls have increased for more transparency by and accountability for those deploying automated decision systems, particularly when those systems impact people's rights. While these two terms are now mentioned regularly in legal and policy debates about AI, it is not always clear what they mean or how they are distinguished from each other.[283] For our purposes, transparency involves measures that provide more and meaningful information about these systems and that, ideally, enable accountability, which involves measures that make companies more responsible for outcomes and impact.[284] That ideal for transparency will not always be attainable, such as when consumers cannot consent to or opt out of corporate use of these systems.

Many proposals exist for how to attain these intertwined goals, which platforms certainly won't reach on their own. These proposals often cover the use of AI tools to address online harms. Below is a brief overview of these goals, with possible legislation discussed later. A major caveat is that even major success on these goals would not actually prevent the harms discussed herein. But it would provide information on the efficacy and impact of these tools, which would help to prevent over-reliance on them, assess whether and when a given tool is appropriate to use, determine the most needed safeguards for such use, and point to the measures the public and private sectors should prioritize to address those harms.[285]

In *Algorithms and Economic Justice*, Commissioner Slaughter identified fairness, transparency, and accountability as the critical principles for systems designed to address algorithmic harms.[286] Meaningful transparency would mean disclosure of intelligible information sufficient to allow third parties to test for discriminatory and harmful outcomes and for consumers to "vote with their feet."[287] Real accountability would mean "that companies—the same ones that benefit from

---

[282] *See* Austin Clyde, *Human-in-the-Loop Systems Are No Panacea for AI Accountability*, Tech Policy Press (Dec. 1, 2021), https://techpolicy.press/human-in-the-loop-systems-are-no-panacea-for-ai-accountability/; Green, *supra* note 280; Green and Kak, *supra* note 278; Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, Engaging Science, Technology, and Society 5 (2019), https://doi.org/10.17351/ests2019.260.

[283] *See, e.g.,* Heidi Tworek and Alicia Wanless, *Time for Transparency From Digital Platforms, But What Does That Really Mean?*, Lawfare (Jan. 20, 2022), https://www.lawfareblog.com/time-transparency-digital-platforms-what-does-really-mean.

[284] *See, e.g.,* Mike Ananny and Kate Crawford, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, New Media and Society 20:3, 973-89 (2018), http://mike.ananny.org/papers/anannyCrawford_seeingWithoutKnowing_2016.pdf.

[285] *See, e.g.*, Shenkman, *supra* note 224 at 35-36; Daphne Keller and Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, in Social Media and Democracy: The State of the Field and Prospects for Reform (Nathan Persily and Joshua A. Tucker, eds.) (Aug. 2020), https://doi.org/10.1017/9781108890960.

[286] Slaughter, *supra* note 13 at 48.

[287] *Id.* at 49. *See also* https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

the advantages and efficiencies of algorithms—must bear the responsibility of (1) conducting regular audits and impact assessments and (2) facilitating appropriate redress for erroneous or unfair algorithmic decisions."[288]

Other government agencies and officials speaking out on these issues include the WHOSTP, which stated that their proposed AI bill of rights might include: "your right to know when and how AI is influencing a decision that affects your civil rights and civil liberties; your freedom from being subjected to AI that hasn't been carefully audited to ensure that it's accurate, unbiased, and has been trained on sufficiently representative data sets; your freedom from pervasive or discriminatory surveillance and monitoring in your home, community, and workplace; and your right to meaningful recourse if the use of an algorithm harms you."[289] Further, the Government Accountability Office stressed that, to build public trust in the use of AI systems, we need independent mechanisms and auditors to "detect error or misuse and ensure equitable treatment of people affected by AI systems."[290] Other executive officials have also promoted transparency and accountability,[291] as have organizations around the globe.[292]

Two fundamental concepts that underly the basic principles and recommendations above are that AI tools, whether or not used for detecting online harms, be both *explainable* and *contestable*. If they lack these features, then those tools are merely "black boxes" not worthy of trust.[293] One government agency, DARPA, engaged in a four-year project regarding the creation of explainable AI (XAI), focused on ensuring that users can understand, trust, and manage these systems.[294] Separately, an interdisciplinary team of academics explored explainability and found that it should often be technically feasible though sometimes practically difficult, and recommended that AI systems be held to the same standard of explainability as humans.[295] In

---

[288] *Id.* at 51.

[289] *See* Lander and Nelson, *supra* note 246.

[290] GAO, *supra* note 43 at 9. *See also* Exec. Order No. 13960, *supra* note 251; OMB Memo, *supra* note 251; NAIRD Strategic Plan, *supra* note 251; Harris, *supra* note 251 at 11, 42-43.

[291] *See* Exec. Order No. 13960, *supra* note 251; OMB Memo, *supra* note 251; NAIRD Strategic Plan, *supra* note 251; Harris, *supra* note 251 at 11, 42-43.

[292] *See, e.g.,* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf.

[293] Distinct from explainability, which often refers to opening black-box machine learning models to understand their decisions after the fact, is the concept of interpretability, which refers to making these models inherently interpretable and thus both easier to understand and less prone to post-hoc manipulation. *See* Cynthia Rudin, et al., *Interpretable machine learning: Fundamental principles and 10 grand challenges*, Statist. Surv. 16: 1-85 (2022), https://doi.org/10.1214/21-SS133. *See also* NIST Special Publication 1270, *supra* note 249 at 26, 38 (noting the import of explainability and interpretability, in part to counteract the view of AI systems "as magic").

[294] *See* https://www.darpa mil/program/explainable-artificial-intelligence; David Gunning, et al., *DARPA's explainable AI (XAI) program: A retrospective*, Applied AI Letters (Dec. 4, 2021), https://doi.org/10.1002/ail2.61.

[295] *See* Finale Doshi-Velez, et al., *Accountability of AI Under the Law: The Role of Explanation*, Berkman Center Research Publication (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761.

contrast, contestability is a design feature by which people "can engage with and challenge" the systems that subject them to automated decisions."[296]

A key document elaborating on all of these themes — and one specifically relevant to addressing online harms — is *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, first issued in 2018 and then revised in 2021.[297] These recommendations, developed by human rights organizations and advocates, are not designed as a regulatory template but reflect "initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights."[298] Among other things, companies should ensure that users know when automated tools are making moderation decisions and should have "a high level understanding" of the decision-making logic.[299] Companies should publicly disclose in regular reports a variety of numbers involving the decisions of these tools, and they should provide certain information — and well-defined appeal rights — to people whose content has been removed or otherwise "actioned."[300] To ensure that the tools are reliable and effective, companies should pursue and monitor detection methods for accuracy and nondiscrimination, submit to regular assessments, and be "encouraged to publicly share data about the accuracy of their systems and to open their process and algorithmic systems to periodic external auditing."[301]

It would also be helpful to create common definitions and standard metrics so that the public and researchers could make cross-platform comparisons.[302] Other recommended disclosures beyond numbers and metrics include explanations of how algorithms are trained and deployed and

---

[296] *See* Daniel N. Kluttz, et al., *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, in After the Digital Tornado: Networks, Algorithms, Humanity at 137-52 (Kevin Werbach, ed.) (2020), https://www.cambridge.org/core/books/after-the-digital-tornado/shaping-our-tools-contestability-as-a-means-to-promote-responsible-algorithmic-decision-making-in-the-professions/311281626ECA50F156A1DDAE7A02CECB.

[297] *See* https://santaclaraprinciples.org/. Other reports make similar recommendations. *See, e.g.*, Future of Tech Commission, *The Future of Tech: A Blueprint for Action* at 18 (2022), https://www.futureoftechcommission.org/; Caitlin Vogus and Emma Llansó, *Making Transparency Meaningful: A Framework for Policymakers*, Center for Democracy and Technology (Dec. 2021) (describing promises and challenges of different types of transparency), https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers/; Singh, *Everything in Moderation*, *supra* note 224 at 33-35; Tech Against Terrorism, *Guidelines on transparency reporting of online counterterrorism efforts*, https://transparency.techagainstterrorism.org. In 2018, Facebook chartered an academic group that made recommendations (some unheeded) on improving the company's public transparency reports. *See* Yale L. S. Justice Collaboratory, *Report of the Facebook Data Transparency Advisory Group* (Apr. 2019), https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

[298] *See* https://santaclaraprinciples.org/?s=03.

[299] *Id.*

[300] *Id. See also* Nicolas P. Suzor, et al., *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, Int'l J. of Communication 13: 1526–1543 (2019), https://ijoc.org/index.php/ijoc/article/view/9736.

[301] *Id.*

[302] *See* Aspen Institute, *supra* note 8 at 20.

platform policies and procedures for borderline rule-breaking content.[303] A separate category of useful data involves the content that platforms have deleted, via automation or humans, but that may be crucial evidence in, e.g., terrorism or war crime cases.[304] The need for such data thus does not center on how platforms made decisions or whether they were correct, and it could be segmented in separate locations with limited access privileges.[305]

Researcher access

As expressed above, platforms should provide not only public reports but also researcher access to data on the use of automated decision tools for potentially harmful content. Researcher access to platform data has received much recent attention. In the aftermath of a Facebook action against New York University researchers, Samuel Levine, Director of the FTC's Bureau of Consumer Protection, stated that the agency "supports efforts to shed light on opaque business practices, especially around surveillance-based advertising," including "good-faith research in the public interest."[306] The Surgeon General has advocated for platforms to "[g]ive researchers access to useful data to properly analyze the spread and impact of misinformation."[307] These calls have been echoed repeatedly in civil society and academia.[308] Beyond just providing data,

---

[303] *See* Centre for Data Ethics and Innovation, *The role of AI in addressing misinformation on social media platforms* at 32-36 (Aug. 5, 2021), https://www.gov.uk/government/publications/the-role-of-ai-in-addressing-misinformation-on-social-media-platforms. *See also* Singh, *Everything in Moderation*, *supra* note 224 at 20.

[304] *See, e.g.*, Tech Against Terrorism, *supra* note 258 at 32, 43; Human Rights Watch, *"Video Unavailable": Social Media Platforms Remove Evidence of War Crimes* (Sep. 10, 2020), https://www hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes; Dia Kayyali, *Human rights defenders are not terrorists, and their content is not propaganda*, WITNESS Blog (Jan. 2, 2020), https://blog.witness.org/2020/01/human-rights-defenders-not-terrorists-content-not-propaganda/; Avi Asher-Shapiro, *YouTube and Facebook Are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals*, The Intercept (Nov. 2, 2017), https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/.

[305] *See* John Bowers, et al., *Digital Platforms Need Poison Cabinets*, Slate (Aug. 24, 2021), https://slate.com/technology/2021/08/social-media-content-moderation-giftschrank.amp?__twitter_impression=true.

[306] *See* https://www.ftc.gov/news-events/blogs/consumer-blog/2021/08/letter-acting-director-bureau-consumer-protection-samuel. *See also* Gilad Edelman, *Facebook's Reason for Banning Researchers Doesn't Hold Up*, WIRED (Aug. 4, 2021), https://www.wired.com/story/facebooks-reason-banning-researchers-doesnt-hold-up/?mc_cid=45f1595320&mc_eid=6ccf77fdd7. Facebook has blocked the work of other researchers, too, as well as failing to give requested data on COVID-19 disinformation to the White House. *See* Elizabeth Dwoskin, et al., *Only Facebook knows the extent of its disinformation problem. And it's not sharing, even with the White House*, The Washington Post (Aug. 19, 2021), https://www.washingtonpost.com/technology/2021/08/19/facebook-data-sharing-struggle/.

[307] *See* Surgeon General, *supra* note 242 at 12.

[308] *See, e.g.*, European Digital Media Observatory and George Washington University Institute for Data, Democracy & Politics, *Report of the Digital Media Observatory's Working Group on Platform-to-Researcher Data Access* (May 31, 2022), https://edmo.eu/2022/05/31/edmo-releases-report-on-researcher-access-to-platform-data/; Renée DiResta, et al., *It's Time to Open the Black Box of Social Media*, Scientific American (Apr. 28, 2022), https://www.scientificamerican.com/article/its-time-to-open-the-black-box-of-social-media/?s=03; Aspen Institute, *supra* note 8 at 20, 28; Singh, *Everything in Moderation*, *supra* note 224 at 34; Susan Benesch, *Nobody Can See Into Facebook*, The Atlantic (Oct. 30, 2021), https://www.theatlantic.com/ideas/archive/2021/10/facebook-oversight-data-independent-research/620557/?s=03; Ethan Zuckerman, *Demand five precepts to aid social-media watchdogs*,

platforms could also allow researchers to perform testing for ecological validity, i.e., real platform users in real-world situations.[309] Such access would allow, e.g., independent analysis of different platform interventions regarding harmful content.[310] Proposed legislation to allow for researcher access is discussed below and may need to wrestle with concerns such as (1) the vetting and protection of researchers, (2) whether investigative journalists or others count as researchers, (3) security and privacy protections for user data,[311] and (4) whether the data was obtained by coercive means, such as the use of dark patterns.

To be clear, it is not that no platforms provide any access to researchers. The issue is that they generally do not provide nearly enough, access is often conditioned on non-disclosure agreements, and some platforms are more open than others. In January 2022, Twitter announced that it is working on privacy-enhancing technology that would allow sharing of more information with researchers, partnering with OpenMined, a non-profit entity.[312] In December 2021, it discussed plans to expand its dataset releases to researchers into areas "including misinformation, coordinated harmful activity, and safety."[313] Other large platforms have either

---

Nature 597, 9 (Aug. 31, 2021), https://doi.org/10.1038/d41586-021-02341-9; Matthias Vermeulen, *The Keys to the Kingdom*, Knight First Amendment Institute (Jul. 27, 2021), https://knightcolumbia.org/content/the-keys-to-the-kingdom. *See also* Harris, *supra* note 251 at 1, 12-13.

[309] *See* Irene V. Pasquetto, et al., *Tackling misinformation: What researchers could do with social media data*, Harvard Kennedy School Misinformation Rev. 1(8) (Dec. 2020), https://doi.org/10.37016/mr-2020-49.

[310] Another way in which expanding researcher access (and public-private cooperation in general) can help achieve meaningful transparency and accountability of relevant AI tools is via examining the extent to which different mechanisms for these ends are working in concert with each other. *See* Spandana Singh and Leila Doty, *Cracking Open the Black Box: Promoting Fairness, Accountability, and Transparency Around High-Risk AI,* New America (Sep. 2021), https://www.newamerica.org/oti/reports/cracking-open-the-black-box/.

[311] The FTC has advised businesses for many years to take privacy and security into account when collecting or using consumers' personal data. Scholars have noted that privacy trade-offs may need to be weighed when considering the value of third-party access to such data, which may derive from people whose information is used to train an AI system or is collected after deployment. *See, e.g.*, *Platform Transparency: Understanding the Impact of Social Media*, S. Comm. on the Judiciary, 117th Cong. (2022) (panelists discussed privacy issues involved in providing access to platform data), https://www.judiciary.senate.gov/meetings/platform-transparency-understanding-the-impact-of-social-media?s=03; Daphne Keller, *User Privacy vs. Platform Transparency: The Conflicts Are Real and We Need to Talk About Them*, Center for Internet and Society Blog (Apr. 6, 2022), https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0?s=03; Sarah Villeneuve, et al., *Shedding Light on the Trade-offs of Using Demographic Data for Algorithmic Fairness*, Partnership on AI (Dec. 2, 2021), https://partnershiponai.org/paper/fairer-algorithmic-decision-making-and-its-consequences/; Hongyan Chang and Reza Shokri, *On the Privacy Risks of Algorithmic Fairness* (Apr. 7, 2021), https://arxiv.org/abs/2011.03731; Nathaniel Persily and Joshua A. Tucker, *Conclusion: The Challenges and Opportunities for a Social Media Research*, in Social Media and Democracy, *supra* note 285 at 313-30; Martin Giles, *The Cambridge Analytica affair reveals Facebook's "Transparency Paradox*," MIT Tech. Rev. (Mar. 20, 2018), https://www.technologyreview.com/2018/03/20/144577/the-cambridge-analytica-affair-reveals-facebooks-transparency-paradox/.

[312] *See* https://blog.twitter.com/engineering/en_us/topics/insights/2022/investing-in-privacy-enhancing-tech-to-advance-transparency-in-ML.

[313] *See* https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed and https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-

not made such pledges or have blocked such access. Of course, there can be too much transparency, in that some data could be incomprehensible, expose sensitive information, or help bad actors figure out how to evade platform policies and filters.

Assessments and audits

Algorithmic Impact Assessments (AIAs) are a means of assessing AI systems in the private or public sector and are derived from assessments performed in environmental protection, human rights, privacy, and data security domains. They allow for the evaluation of an AI system's impact before, during, or after its use. Further, they can allow companies to mitigate bad outcomes and, if publicly shared, provide a chance for accountability and safer, better use of the technology.[314] AIAs could also provide the FTC and other regulators with information for investigations into deceptive and unfair business practices. The need for AIAs is recognized broadly, including for content moderation,[315] and many frameworks for implementing them have been proposed, both here and abroad.[316] Legislative attention to them is discussed later.

Major questions for developing these assessments include when they should be conducted, which entities should be subject to them, and whether they should be performed internally or via external auditors. Another fundamental determination is whether such an assessment is conceived as an AIA or as an audit.[317] Although sometimes AIA and audit are used interchangeably, the former may refer more often to a focus on algorithmic design, possible harm, and ultimate responsibility, whereas an audit may refer more often to evaluation of an AI

---

operations-; *See also* Camille Francois, *The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations*, J. of Online Trust and Safety (Oct. 2021), https://tsjournal.org/index.php/jots/article/view/17/8.

[314] *See, e.g.*, *H. Comm. on Science, Space, and Technology* (testimony of Meredith Whittaker), *supra* note 252; *Comments of AI NOW Institute*, FTC Hearings on Competition and Consumer Protection in the 21st Century (Aug. 20, 2018), https://ainowinstitute.org/ainow-ftc-comments-consumer-protection.pdf.

[315] *See, e.g.*, Aspen Institute, *supra* note 8 at 21, 37; United Nations Special Rapporteur, *Report on Artificial Intelligence technologies and implications for freedom of expression and the information environment* (Aug. 29, 2018), https://www.undocs.org/A/73/348; O'Neil, *supra* note 235, at 207, 217; Sylvain, *Recovering Tech's Humanity*, *supra* note 229 at 281.

[316] An October 2021 House hearing focused on AI ethics and transparency, with several witnesses discussing the need for audits and AIAs and referring to proposed and existing frameworks, particularly for bias detection. *See Task Force on Artificial Intelligence: Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age*, H. Comm. on Financial Services, 117th Cong. (2021), https://www.congress.gov/event/117th-congress/house-event/114125. *See also* Data & Society, *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest* (Jun. 29, 2021), https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/; Ada Lovelace Institute, *Technical methods for regulatory inspection of algorithmic systems* (Dec. 9, 2021), https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/; United Kingdom Government Digital Service, *Data Ethics Framework* (2020), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923108/Data_Ethics_Framework_2020.pdf.

[317] *See* Khari Johnson, *The Movement to Hold AI Accountable Gains More Steam*, WIRED (Dec. 2, 2021), https://www.wired.com/story/movement-hold-ai-accountable-gains-steam/?s=03.

model's output.[318] Further, to ensure that these assessments are meaningful and comparable, standards must be set for how AIAs or audits should be conducted and what they should include.[319] Similarly, the results of any such assessments need to be documented in some standardized way.[320] Recognized standards and documentation would also help to allow for better evaluation of an auditor's work. Again, assessments and audits are important but not a substitute for enforcement and remedies relating to online harms.

Auditor and employee protections

AIAs and algorithmic audits may also not be successful if the auditors doing the work are not certified, independent, and protected in their work.[321] These concerns may well increase as the small marketplace of outside auditors grows.[322] Workers within tech companies need protection, too, when they seek to report on harm or unfairness that AI tools are facilitating or failing to block, whether in the role of a whistleblower or otherwise.[323] In the words of Timnit Gebru, "[t]he baseline is labor protection and whistleblower protection and anti-discrimination laws.

---

[318] *See id.*; Danaë Metaxa, et al., *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In*, Foundations and Trends in Human–Computer Interaction 14:4 (2021), http://dx.doi.org/10.1561/1100000083.

[319] *See, e.g.*, UK Digital Regulation Cooperation Forum, *Auditing algorithms: the existing landscape, role of regulators and future outlook* (Apr. 28, 2022), https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook; Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 Harvard J. of Law & Tech. __ (forthcoming) (Jun. 15, 2021), https://ssrn.com/abstract=3867634; Gregory Falco, et al., *Governing AI safety through independent audits*, Nature Machine Intelligence 3: 566-71 (2021), https://www.nature.com/articles/s42256-021-00370-7; Aspen Institute, *supra* note 8 at 20; Johnson, *supra* note 317; Moana Slone, *The Algorithmic Auditing Trap*, OneZero (Mar. 17, 2021), https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d; Hayden Field, *Seven AI ethics experts predict 2022's opportunities and challenges for the field*, Morning Brew (Jan. 17, 2022) (quoting Deborah Raji and Abhishek Gupta), https://www.morningbrew.com/emerging-tech/stories/2022/01/17/seven-ai-ethics-experts-predict-2022-s-opportunities-and-challenges-for-the-field?s%E2%80%A6; Kate Kaye, *A new wave of AI auditing startups wants to prove responsibility can be profitable*, Protocol (Jan. 3, 2022), https://www.protocol.com/enterprise/ai-audit-2022?s=03#toggle-gdpr.

[320] *See, e.g.*, Selbst, *supra* note 319; Inioluwa Deborah Raji, et al., *Closing the Accountability Gap, Defining an End-to-End Framework for Internal Algorithmic Auditing*, in Conf. on Fairness, Accountability, and Transparency (FAT '20) (Jan. 2020), https://doi.org/10.1145/3351095.

[321] *See* https://hai.stanford.edu/news/radical-proposal-third-party-auditor-access-ai-accountability?s=03. *See also* NIST Special Publication 1270, *supra* note 249 at 36 (noting concern that technology companies being assessed not "have undue influence on building or using the assessment"); J. Nathan Mathias, *Why We Need Industry-Independent Research on Tech & Society*, CAT Lab (Jan. 2020) (discussing research management of conflicts of interest), https://citizensandtech.org/2020/01/industry-independent-research/.

[322] *See* Khari Johnson, *What algorithm auditing startups need to succeed*, VentureBeat (Jan. 30, 2021), https://venturebeat.com/2021/01/30/what-algorithm-auditing-startups-need-to-succeed/.

[323] *See, e.g.*, Erie Meyer, *CFPB Calls Tech Workers to Action*, At the CFPB Blog (Dec. 15, 2021), https://www.consumerfinance.gov/about-us/blog/cfpb-calls-tech-workers-to-action/?s=03; Brennan Center, *supra* note 257 at 3; *H. Comm. on Science, Space, and Technology* (testimony of Meredith Whittaker), *supra* note 252.

Anything we do without that kind of protection is fundamentally going to be superficial, because the moment you push a little bit, the company's going to come down hard."[324]

Other considerations

Other proposals to increase transparency and accountability for AI tools run the gamut from "bug bounties," which are designed to bring out hidden biases,[325] to independent bodies such as the Facebook Oversight Board, made up of 20 outside experts from around the world who consider appeals of Facebook and Instagram content decisions.[326] Some of the Board's recommendations have touched on automated removal of TVEC and the need for the company to be more transparent about such removal decisions and to provide better notice and appeal rights to users.[327]

A crucial issue behind calls for increased disclosure of data is how to do so while maintaining user privacy.[328] Deidentification of such data is one theoretical way to address it,[329] as is user consent, though practical and meaningful ways to obtain them may be challenging if not impossible.[330] The use of differential privacy and synthetic data are other potential solutions, though not ones without any risk of data leakage.[331] The UN has recognized privacy risks while

---

[324] Dina Bass, *Google's Former AI Ethics Chief Has a Plan to Rethink Big Tech*, Bloomberg Businessweek (Sep. 20, 2021), https://www.bloomberg.com/news/articles/2021-09-20/timnit-gebru-former-google-ai-ethics-chief-has-plan-to-rethink-big-tech.

[325] *See, e.g.,* Algorithmic Justice League, *Bug Bounties for Algorithmic Harms?* (Jan. 2022), https://www.ajl.org/bugs; Rumman Chowdhury and Jutta Williams, *Introducing Twitter's first algorithmic bias bounty challenge*, Twitter Engineering Blog (Jul. 30, 2021), https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge; Khari Johnson, *AI researchers propose 'bias bounties' to put ethics principles into practice*, VentureBeat (Apr. 17, 2020), https://venturebeat.com/2020/04/17/ai-researchers-propose-bias-bounties-to-put-ethics-principles-into-practice/.

[326] See https://oversightboard.com/.

[327] *See* Dia Kayyali and Jillian C. York, *The Facebook Oversight Board is making good decisions- but does it matter?*, Tech Policy Press (Jul. 28, 2021), https://techpolicy.press/the-facebook-oversight-board-is-making-good-decisions-but-does-it-matter/.

[328] *See, e.g.,* Aspen Institute, *supra* note 8 at 21-22, 28.

[329] *See* Surgeon General, *supra* note 242 at 12.

[330] . *See, e.g.*, Katherine Miller, *De-Identifying Medical Patient Data Doesn't Protect Our Privacy*, Stanford HAI News (Jul. 19, 2021), https://hai.stanford.edu/news/de-identifying-medical-patient-data-doesnt-protect-our-privacy; Gina Kolata, *Your Data Were 'Anonymized'? These Scientists Can Still Identify You*, The New York Times (Jul. 23, 2019), https://www nytimes.com/2019/07/23/health/data-privacy-protection html. Deidentification could also make it harder to determine how representative a dataset is. Further, even if one could obtain meaningful user consent, it is possible that the choices of certain users to consent or opt out would affect the representativeness of a dataset.

[331] *See* Nathan Persily, *A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act* at 4, J. of Online Trust and Safety 1:1 (Oct. 28, 2021) (arguing that laws allowing for research access to platform data should ensure anonymity and encouraging use of differential privacy and construction of synthetic datasets), https://doi.org/10.54501/jots.v1i1.22. *See also* Joseph Near and David Darais, *Differentially Private Synthetic Data*, NIST Cybersecurity Insights (May 3, 2021) (considering value of differentially private synthetic data), https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data#; Meg

still supporting explainable and transparent AI systems, including AIAs and grievance mechanisms.[332] Privacy is also one aspect of trustworthy AI that NIST will study and incorporate into a Congressionally mandated Risk Management Framework.[333]

The many challenges of transparency and accountability — and the fact that they don't by themselves prevent harm — highlight the importance of focusing on the entire AI lifecycle, design through implementation. Some scholars have argued that transparency might be less important if algorithms could be designed not to discriminate in the first place.[334] Both designers and users of AI tools must nonetheless continue to monitor the impact of their AI tools, since fair design does not guarantee fair outcomes. In its Online Harms White Paper, the United Kingdom's government indicated it would work with industry and civil society to develop a Safety by Design framework for online services, possibly to include guidance on effective systems for addressing illegal or harmful content via AI and trained moderators.[335] Algorithmic design is not within the scope of this report, though it is referred to again in the discussion below on platform interventions.

D. Responsible data science

Those building AI systems, including tools to combat online harms, should take responsibility for both inputs and outputs. Such responsibility includes the need to avoid unintentionally biased or unfair results derived from problems with the training data, classifications, or algorithmic design. In their call for an AI bill of rights, WHOSTP officials note that some AI failings that disproportionately affect already marginalized groups "often result from AI developers not using appropriate data sets and not auditing systems comprehensively, as well as not having diverse perspectives around the table to anticipate and fix problems before products are used (or to kill products that can't be fixed)."[336] Further, the 2021 DHS report on deepfakes stated that scientists

---

Young, et al., *Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing*, Proc. of ACM (FAT'19) (Jan. 29, 2019) (advocating for use of synthetic data and a third-party public-private data trust), https://par nsf.gov/servlets/purl/10111608.

[332] United Nations High Commissioner for Human Rights (UNHCHR), The right to privacy in the digital age (Sep. 13, 2021), https://www.ohchr.org/EN/Issues/DigitalAge/Pages/cfi-digital-age.aspx.

[333] *See* National Defense Authorization Act for Fiscal Year 2021, H.R. 116-617, § 5301, at 2768-2775, https://www.congress.gov/congressional-report/116th-congress/house-report/617/1?overview=closed; *See also* https://hai.stanford.edu/policy/policy-resources/summary-ai-provisions-national-defense-authorization-act-2021.

[334] *See, e.g.,* Joshua A. Kroll, et al., *Accountable Algorithms*, 165 U. Penn. L. Rev. 633 (2017), https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/.

[335] *See* United Kingdom Department for Digital, Culture, Media, and Sport, and Home Office, *Online Harms White Paper* at 8.14 (Dec. 15, 2020), https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper. The White Paper informed the pending Online Safety Bill, first introduced in May 2021. *See* https://www.gov.uk/government/publications/draft-online-safety-bill.

[336] Lander and Nelson, *supra* note __.246. *See also* NIST Special Publication 1270, *supra* note 249 at 36-37, 45 (noting benefits of diversity within teams training and deploying AI systems, that "the AI field noticeably lacks diversity," and that team supervisors should be responsible for risks and associated harms of these systems); Color of Change, *Beyond the Statement: Tech Framework* (also recommending that decision-makers be held responsible for discriminatory outcomes), https://beyondthestatement.com/tech-framework/.

should be considering at the development stage how to mitigate potential misuses of deepfake models.[337]

Developers who fund, oversee, or direct scientific research in this area should appreciate that their work does not happen in a vacuum and address the fact that it could cause harm.[338] This recognition includes the fundamental idea that the data being used has context and often stands in for real people.[339] To move individual scientists in this direction, a large AI conference instituted a requirement that submitting authors include a statement on the broader societal impacts of their research.[340] The Partnership on AI has issued recommendations on how those leading and directing research can anticipate and mitigate any potential negative impacts.[341] One important and practical consideration is having adequate documentation throughout the AI development process.[342] Further, various scholars have proposed reparative approaches to AI development and redress.[343]

Unconscious bias of researchers, or at least a failure to actively consider bias and its mitigation, can also create problems.[344] The MIT Media Lab created a system to help researchers deal with unconscious biases at different stages of a project.[345] A broader and more significant solution is to deal with the lack of diversity in the AI field, including in the ranks of decision-makers as well as in the research teams working on these matters.[346] The inclusion of diverse viewpoints should

---

[337] *See* DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 31.

[338] *See* O'Neil, *supra* note 235 at 205 ("Like doctors, data scientists should pledge a Hippocratic Oath, one that focuses on the possible misuses and misinterpretations of their models."); H. Comm. on Science, Space and Technology (testimony of Joy Buolamwini), *supra* note __.252. Within technology companies, putting the burden of raising or addressing these issues solely on employees, rather than their superiors, can put employees in an untenable position of deciding whether doing the right thing is worth the risk that they could lose their jobs for doing so. *See* Inga Strümke, et al., *The social dilemma in artificial intelligence development and why we have to solve it*, (Dec. 2021) (arguing for creation of professional AI code of ethics), https://doi.org/10.1007/s43681-021-00120-w.

[339] *See* Inioluwa Deborah Raji, *The Discomfort of Death Counts: Mourning through the Distorted Lens of Reported COVID-19 Death Data,* Patterns (N Y) 1(4): 100066 (Jul. 10, 2020), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7296309/.

[340] *See* https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/; Carina E. A. Prunkl, et al., *Institutionalizing ethics in AI through broader impact requirements*, Nature Machine Intelligence 3, 104-110 (2021), https://www nature.com/articles/s42256-021-00298-y.

[341] *See* Partnership on AI, *Managing the Risks of AI Research: Six Recommendations for Responsible Publication* (May 6, 2021), https://partnershiponai.org/paper/responsible-publication-recommendations/.

[342] *See* NIST Special Publication 1270, *supra* note 249 at 44; Selbst, *An Institutional View of Algorithmic Impact Assessments*, *supra* note 319; Raji, *Closing the AI Accountability Gap, supra* note 320 at 37.

[343] *See* Khari Johnson, *A Move for 'Algorithmic Reparation' Calls for Racial Justice in AI*, WIRED (Dec. 23, 2021), https://www.wired.com/story/move-algorithmic-reparation-calls-racial-justice-ai/.

[344] *See* Deborah Raji, *How our data encodes systematic racism*, MIT Tech. Rev. (Dec. 10, 2020), https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/.

[345] *See* https://aiblindspot media mit.edu/.

[346] Several witnesses discussed the importance of researcher diversity in 2021 and 2019 House hearings. *See Task Force on Artificial Intelligence*, *supra* note 316 (testimony of Miriam Vogel and Aaron Cooper); H. Comm. on

be meaningful and include people with decision-making authority; it should not be used to engage in what amounts to "participation-washing."[347] To get such viewpoints, a strong pipeline of people need to be trained and hired for these roles, something that groups like Black in AI, Queer in AI, and LatinX in AI are working to achieve. Firms need to retain such people, once hired, by striving to create and maintain diverse, equitable, inclusive, and accessible cultures in which such people no longer face marginalization, discrimination, or exclusion.[348] Of course, the composition of the team designing an AI model does not necessarily alter discriminatory or biased outcomes of that model if they stem from problems in the underlying data.[349]

An even larger issue is that only a few big technology companies are funding most of the research in question.[350] They are also able to capture, within their companies or academia, institutions or researchers who may then be likely to work in accord with corporate aims.[351] They can also use their dominant positions and wealth to set the agenda for what AI research the government will and will not fund, again in line with their own incentives.[352] Some prominent researchers, like Timnit Gebru, have started their own AI research centers to deal with these

---

Science, Space and Technology, *supra* note 252 (testimony of Meredith Whittaker and Joy Buolamwini), *supra* note __.252. *See also* UNHCHR, *supra* note 332 at 16; Sue Shellenbarger, *A Crucial Step for Averting AI Disasters*, The Wall St. J. (Feb. 13, 2019), https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865; Sasha Costanza-Chock, *Design Justice: towards an intersectional feminist framework for design theory and practice*, Proc. of the Design Research Society 2018 (Jun. 3, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3189696.

[347] *See* Mona Sloane, *Here's what's missing in the quest to make AI fair*, Nature (May 5, 2022), https://www.nature.com/articles/d41586-022-01202-3; Mona Sloane, *Participation-washing could be the next dangerous fad in machine learning*, MIT Tech. Rev. (Aug. 25, 2020), https://www.technologyreview.com/2020/08/25/1007589/participation-washing-ai-trends-opinion-machine-learning/. *See also* NIST Special Publication 1270, *supra* note 249 at 36 (suggesting that different kinds of diversity are important to consider in terms of power and decision-making within technology companies).

[348] *See, e.g.,* Dr. Jeffrey Brown, *After the Offer: The Role of Attrition in AI's 'Diversity Problem*,' Partnership on AI (Apr. 13, 2022), https://partnershiponai.org/paper/after-the-offer-the-role-of-attrition-in-ais-diversity-problem/; Megan Rose Dickey, *Examining the "pipeline problem,"* TechCrunch (Feb. 14, 2021), https://techcrunch.com/2021/02/14/examining-the-pipeline-problem/; *Inclusion in Tech: How Diversity Benefits All Americans*, H. Comm. on Energy and Commerce, 116th Cong. (2019) (testimony of Nicol Turner-Lee), https://docs.house.gov/meetings/IF/IF17/20190306/108901/HHRG-116-IF17-Wstate-Turner-LeeN-20190306.pdf.

[349] *See* Benjamin, *supra* note 245 at 59.

[350] *See* Karen Hao, *Inside the fight to reclaim AI from big tech's control*, MIT Tech. Rev. (Jun. 14, 2021), https://www.technologyreview.com/2021/06/14/1026148/ai-big-tech-timnit-gebru-paper-ethics/; *House Comm. on Science, Space, and Technology*, *supra* note 252 (testimony of Meredith Whittaker), *supra* note 252.

[351] *See* Meredith Whittaker, *The Steep Cost of Capture*, Interactions 28(6), 50 (Nov.-Dec. 2021), https://interactions.acm.org/archive/view/november-december-2021/the-steep-cost-of-capture; Abeba Birhane, et al., *The Values Encoded in Machine Learning Research* (Jun. 2021), https://arxiv.org/pdf/2106.15590.pdf.

[352] *See* Timnit Gebru, *For truly ethical AI, its research must be independent from big tech*, The Guardian (Dec. 6, 2021), https://www.theguardian.com/commentisfree/2021/dec/06/google-silicon-valley-ai-timnit-gebru?s=03; Laurie Clarke, et al., *How Google quietly funds Europe's leading tech policy institutes*, The New Statesman (Jul. 30, 2021), https://www newstatesman.com/science-tech/big-tech/2021/07/how-google-quietly-funds-europe-s-leading-tech-policy-institutes.

problems; hers will focus on harm to marginalized groups.[353] Congress has asked about how to foster innovative ways to combat online harm, and thus one response, in her words, is that "what truly stifles innovation is the current arrangement where a few people build harmful technology and others constantly work to prevent harm, unable to find the time, space or resources to implement their own vision of the future."[354]

Finally, it is critical that the research community keep privacy in mind. AI development often involves huge amounts of training data, which can be amassed in invasive ways[355] and which is in tension with data minimization principles. As noted above in the transparency context, implementing adequate privacy protections for such data may be difficult in practice and may require creative solutions. Eventually, AI systems may be trained on much less data, as opposed to the current hunger for more, but it is unclear how long it may take for that to happen.[356]

E. Platform AI interventions

Mitigation tools

The use of automated tools to address online harms is most often framed as an issue of detection and removal, whether before or after content is posted. But platforms, search engines, and other technology companies can and do use these tools to address harmful content in other ways. They have a range of interventions or "frictions" to employ, including circuit-breaking, downranking, labeling, adding interstitials, sending warnings, and demonetizing bad actors.[357] Some platforms already use such mitigation measures, but their relative secrecy means few details are known at either a systemic or individual level about their efficacy or impact. These interventions are generally automated and thus many of them would have the same inherent flaws of AI-based detection tools, as they would still be dependent on the ability to identify particular types of

---

[353] *See, e.g.*, Nitasha Tiku, *Google fired its star AI researcher one year ago. Now she's launching her own institute*, The Washington Post (Dec. 2, 2021), https://www.washingtonpost.com/technology/2021/12/02/timnit-gebru-dair/?s=03; Tom Simonite, *Ex-Googler Timnit Gebru Starts Her Own AI Research Center*, WIRED (Dec. 2, 2021), https://www.wired.com/story/ex-googler-timnit-gebru-starts-ai-research-center/?s=03.

[354] Gebru, *supra* note 352.

[355] *See, e.g.*, John McQuaid, *Limits to Growth: Can AI's Voracious Appetite for Data BeTamed?*, Undark (Oct. 18, 2021), https://undark.org/2021/10/18/computer-scientists-try-to-sidestep-ai-data-dilemma/.

[356] *See, e.g.*, Tom Simonite, *Facebook Says Its New AI Can Identify More Problems Faster*, WIRED (Dec. 8, 2021), https://www.wired.com/story/facebook-says-new-ai-identify-more-problems-faster/; H. James Wilson, et al., *The Future of AI Will Be About Less Data, Not More*, Harvard Bus. Rev. (Jan. 14, 2019), https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more.

[357] One proffered example of demonetization is for Google to use probability scores that AI assigns to violative content in search results, which results in its blocking or demotion, to penalize misinformation-filled sites "in the algorithmic auctions Google runs in which sites … bid for ad placements." Noah Giansiracusa, *Google Needs to Defund Misinformation*, Slate (Nov. 18, 2021), https://slate.com/technology/2021/11/google-ads-misinformation-defunding-artificial-intelligence html. *See also* Ryan Mac, *Buffalo gunman's video is surfacing on Facebook, sometimes with ads beside it*, The New York Times (May 19, 2022), https://www.nytimes.com/2022/05/19/technology/buffalo-shooting-facebook-ads html.

potentially harmful content.[358] As noted above, such content may need detection only because platform recommendation engines, powered by AI, can spread and amplify it so well. In any event, these interventions need more study, which means much more transparency about their use and effects, with due access for research.

Several major reports on online harm and disinformation discuss the potential value, pitfalls, and lack of transparency regarding various platform interventions, including reports from the Aspen Institute, the Brennan Center for Justice, the Global Disinformation Index, the Coalition to Fight Digital Deception, and the United Kingdom's Royal Society.[359] In addition, the Surgeon General's advisory on health disinformation states that platforms should build in "frictions" such as suggestions, warnings, and early detection of viral content.[360] They are also discussed at length in several academic papers.[361]

One measure gaining traction since its introduction by Rutgers University Professor Ellen P. Goodman is the use of so-called circuit breakers or virality disruptors.[362] This intervention involves platforms' disrupting traffic at a certain threshold of circulation, at which point human reviewers would assess the content to ensure it does not violate the law or platform policy.[363] Doing so could reduce the fast spread of harmful content and is akin to the steps that stock exchanges can take to curb trading volatility.[364] While one benefit of this intervention is that it does *not* require content-based detection, some have proposed that AI could help determine what viral content should be slowed.[365]

---

[358] For example, a European study on deepfakes called for both content providers and content creators to label deepfakes but noted that deepfake detection software might be a prerequisite for such a requirement. *See* EPRS, *Tackling deepfakes in European policy supra* note 43 at 59-62.

[359] *See* Aspen Institute, *supra* note 8, at 66-67; Brennan Center for Justice, *supra* note 257, at 12-15; CFDD, *supra* note 224 at 16-19; Global Disinformation Index, *Disrupting Online Harms: A New Approach* at 14 (Jul. 2021), www.disinformationindex.org; The Royal Society, *The online information environment: Understanding how the internet shapes people's engagement with scientific information* 13-15 (Jan. 2022), https://www.royalsociety.org/online-information-environment.

[360] Surgeon General, *supra* note 242 at 12.

[361] *See, e.g.*, Eric Goldman, *Content Moderation Remedies*, Santa Clara U. Legal Studies Research Paper (Mar. 24, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810580; Molly K. Land and Rebecca J. Hamilton, *Beyond Takedown: Expanding the Toolkit for Responding to Online Hate*, American U., WCL Research Paper No. 2020-11 (Jan. 31, 2020), https://ssrn.com/abstract=3514234.

[362] *See* Ellen P. Goodman, *Digital Information Fidelity and Friction*, Knight First Amendment Institute (Feb. 26, 2020), https://knightcolumbia.org/content/digital-fidelity-and-friction.

[363] *See id.*

[364] *See id. See also* Center for American Progress, *Fighting Coronavirus Misinformation and Disinformation* at 27-28 (Aug. 2020), https://www.americanprogress.org/article/fighting-coronavirus-misinformation-disinformation/.

[365] *See, e.g.*, Future of Tech Commission, *supra* note 297 at 18; Johns Hopkins University, Imperial College of London, and Georgia Institute of Technology, *Countering disinformation: improving the Alliance's digital resilience*, NATO Review (Aug. 12, 2021), https://www.nato.int/docu/review/articles/2021/08/12/countering-disinformation-improving-the-alliances-digital-resilience/index.html; Young, *supra* note 130 at 5; Christina

Another set of measures subject to much discussion involves contextual labeling, interstitials, user prompts, and warnings. The point of these interventions — which would more likely be dependent on automated, content-based detection tools — is to advise or warn users about what they are about to see or post. In other words, they could protect users from potentially harmful content already circulating or reduce the chance that a given user will publish such content.

Contextual labeling is a familiar concept to the FTC. In the context of false advertising, the FTC often requires in its orders that companies avoid deception by disclosing certain facts, clearly and conspicuously, in close proximity to certain representations the companies make about their products. Assuming compliance, the value of such disclosures depends largely on whether consumers see and understand them. This is also true for contextual labeling of online content.

One recent study reviews the nascent literature on the efficacy of online content labeling, concluding that it shows promise for helping to correct or limit the impact of misinformation but that certain psychological phenomena are at play.[366] The authors are much more concerned about, for example, the "implied truth effect," whereby people believe that unlabeled content must be truthful, than the "backfire effect," whereby people solidify their beliefs in the opposite of whatever such labels tell them. Other recent studies conclude that interstitial warnings — which appear as separate pages or pop-ups that users cannot miss and must take some action to get past — are more effective that contextual ones.[367] It may be that the greater efficacy has more to do with the fact that they are a source of friction than with the particular information provided.[368]

A related type of intervention is a user prompt or warning — which could be in interstitial form — that appears before a user posts or shares potentially harmful content. Several articles and studies have advocated for platforms to use them more often.[369] One study involves the use of

---

Pazzanese, *How the government can support a free press and cut disinformation* (Q&A with Martha Minow) (Aug. 11, 2021), https://news harvard.edu/gazette/story/2021/08/martha-minow-looks-at-ways-government-can-stop-disinformation/.

[366] *See* Garrett Morrow, et al., *The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation* (Dec. 3, 2020), http://dx.doi.org/10.2139/ssrn.3742120. *See also* Emily Saltz, et al., *Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions*, Partnership on AI (Dec. 2020), https://arxiv.org/abs/2011.12758.

[367] *See* Ben Kaiser, et al., *Adapting Security Warnings to Counter Online Disinformation* (Aug. 2020), https://www.usenix.org/system/files/sec21summer_kaiser.pdf; Filipo Sharevski, et al., *Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes* (Apr. 1, 2021), https://arxiv.org/abs/2104.00779.

[368] *See* Kaiser, *supra* note 367 at 14.

[369] *See* Christopher Paul and Hilary Reininger, *Platforms Should Use Algorithms to Help Users Help Themselves*, Carnegie Endowment for International Peace (Jul. 20, 2021), https://carnegieendowment.org/2021/07/20/platforms-should-use-algorithms-to-help-users-help-themselves-pub-84994; Ziv Epstein, et al., *Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online*, Harvard Kennedy School Misinformation Rev. (May 18, 2021), https://misinforeview.hks.harvard.edu/article/developing-an-accuracy-prompt-toolkit-to-reduce-covid-19-misinformation-online/; Gordon Pennycook, et al., *Shifting attention to accuracy can reduce misinformation online*,

machine learning to warn users about sharing harmful content in encrypted messaging apps while avoiding privacy and security issues in that context.[370] While some have noted that these measures may avoid censorship concerns, of course they do not block harmful information being spread by people (or bots) with malicious intent.

Many platforms already use or have experimented with interventions beyond blocking and removing content or suspending accounts. In September 2021, Facebook disclosed its Content Distribution Guidelines, which described types of content that it demotes and the rationales for doing so.[371] Such content includes posts with fact-checked and debunked information, with predicted but not confirmed policy violations (e.g., use of hate terms, graphic violence, or fake accounts) and with suspicious virality.[372] Facebook had first announced it would institute policies for borderline content in 2018.[373] In August 2021, its Instagram property announced it would show stronger warnings when users are about to post potentially offensive or harassing content.[374] Other platforms that have indicated some use of such interventions include YouTube, which demotes some borderline content, and WhatsApp, which places a limit on the number of times content can be forwarded.[375] Further, Nextdoor uses a "Kindness Reminder," an interstitial that runs on machine learning, when a user is about to post something potentially harmful.[376]

The most outspoken platform in this space is likely Twitter. In a set of "open internet" principles, the company states that "content moderation is now more than just leaving content up or taking it down. Providing users with context, whether concerning an account, piece of content, or form of engagement, is more informative to the broader public conversation than removing content while providing controls to people and communities to control their own experience is empowering and impactful. Equally, deamplification allows a more nuanced approach to types of speech that may be considered problematic, better striking a balance between freedom of speech and

---

Nature 592, 590 (2021), https://www.nature.com/articles/s41586-021-03344-2.pdf; Andrew Myers, *The best way to counter fake news is to limit person-to-person spread, Stanford study finds*, Stanford News Service (Oct 25, 2021), https://news.stanford.edu/press-releases/2021/10/25/foil-fake-news-fs-infectiousness/; Mustafa Mikdat Yildirim, et al., *Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter*, Cambridge U. Press (Nov. 22, 2021), https://doi.org/10.1017/S1537592721002589.

[370] *See* Yiqing Hua, *New technology may bridge privacy debate on encrypted messaging*, Tech Policy Press (Oct. 21, 2021), https://techpolicy.press/new-technology-may-bridge-privacy-debate-on-encrypted-messaging/.

[371] *See* https://about fb.com/news/2021/09/content-distribution-guidelines/; https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/.

[372] *Id.; see* Jeff Allan, *The Integrity Institute's Analysis of Facebook's Widely Viewed Content Report [2021-Q4]*, The Integrity Institute (Mar. 30, 2022) (finding that Facebook had failed to block or intervene on popular content failing basic media literacy checks), https://integrityinstitute.org/widely-viewed-content-analysis-tracking-dashboard#other-platforms.

[373] Josh Constine, *Facebook will change algorithm to demote "borderline content" that almost violates policies*, TechCrunch (Nov. 15, 2018), https://techcrunch.com/2018/11/15/facebook-borderline-content/.

[374] *See* https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse.

[375] *See* https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/; https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en.

[376] *See* https://blog.nextdoor.com/2019/09/18/announcing-our-new-feature-to-promote-kindness-in-neighborhoods/.

freedom of reach. Long term, how attention is directed is a critical question."[377] Among other things, Twitter prompts users who try to retweet before reading a post or who are about to send a reply using potentially harmful or abusive text, and it adds labels to tweets with potentially misleading information.[378]

Some platforms and search engines also use external fact-checking organizations to determine whether to intervene with respect to potentially harmful or false materials, including election-related content or TVEC. [379] This approach has its detractors, who may point to questions about norms and standards for these organizations, whether they can scale, their impact, and the lack of transparency surrounding their use. [380] It also has proponents, who argue, for example, that platforms can leverage a robust ecosystem of international websites to make better determinations about what content to downrank or label.[381] Most would agree that it needs more study.[382]

---

[377] *See* Twitter, *Protecting the Open Internet*, https://cdn.cms-twdigitalassets.com/content/dam/about-twitter/en/our-priorities/open-internet.pdf. The reference to "freedom of reach" and its distinction from freedom of speech comes from Renée Diresta. *See* Renée Diresta, *Free Speech Is Not the Same as Free Reach*, WIRED (Aug. 30, 2018), https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/.

[378] *See* https://perma.cc/XM9V-2S8E; https://twitter.com/TwitterSupport/status/1460715806401122305. A recent experiment in which Twitter users were prompted to reconsider before posting potentially offensive content showed some efficacy, including a decrease in later offensive posts from same users. *See* Matthew Katsaros, et al., *Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content*, Int'l AAAI Conf. on Web and Social Media (2022) (also describing the researchers' use of an algorithm that relied on a large language model to determine when to intervene), https://arxiv.org/abs/2112.00773. A recent study on Twitter's election misinformation warning labels indicate that they did not impact engagement with tweeted misinformation but can be more helpful when the label provides a strong rebuttal and uses text similar to the words in the tweet. *See* Orestis Papakyriakopoulos and Ellen P. Goodman, *The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets*, ACM WWW '22 (forthcoming) (February 22, 2022), https://ssrn.com/abstract=4036042.

[379] *See, e.g.*, https://www facebook.com/business/help/2593586717571940?id=673052479947730; https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter; https://blog.google/products/news/fact-checking-misinformation-google-features/; https://blogs.bing.com/Webmaster-Blog/September-2017/Bing-adds-Fact-Check-label-in-SERP-to-support-the-ClaimReview-markup.

[380] *See, e.g.*, DHS Analytic Exchange Program, *Combatting Targeted Disinformation Campaigns, Part Two* at 25-30 (Aug. 2021) (concluding that the impact of fact-checking may be limited but that the efficacy of different methods should be studied), https://www.dhs.gov/publication/2021-aep-deliverables; CFDD, *supra* note 224 at 14 (questioning whether fact-checking efforts can scale and noting that platforms lack transparency about them and apply them inconsistently).

[381] *See, e.g.*, *An open letter to YouTube's CEO from the world's fact-checkers* (Jan. 12, 2022), https://maldita.es/uploads/public/docs/youtube_open_letter_en.pdf; Barrett, *Who Moderates the Social Media Giants?*, *supra* note 281 at 23, 26; Jan Oledan, et al., *Fact-checking networks fight coronavirus infodemic*, Bulletin of the Atomic Scientists (Jun. 25, 2020), https://thebulletin.org/2020/06/fact-checking-networks-fight-coronavirus-infodemic/; Marcus Bösch and Becca Ricks, *Broken Promises: TikTok and the German Election*, Mozilla Foundation (Sep. 2021), https://foundation mozilla.org/en/campaigns/tiktok-german-election-2021/.

[382] *See, e.g.*, NATO Strategic Communications Centre of Excellence, *Inoculation Theory and Misinformation* (Oct. 2021), https://stratcomcoe.org/publications/inoculation-theory-and-misinformation/217; Mohan, *supra* note 234 (discussing YouTube's ongoing consideration of fact-checking and labeling).

Sometimes these fact-checking efforts intersect with AI tools, such as via the use of such tools to debunk false claims,[383] determine what content to send to fact checkers for review,[384] develop datasets of debunked information,[385] help match claims across encrypted messages,[386] or develop chatbots, like one developed by a Spanish fact-checking organization to help WhatsApp users get answers to the veracity of information.[387] Others have discussed the potential use of crowdsourcing, rather than professional fact-checkers, noting that machine learning can facilitate such efforts.[388] At least two vendors, Logically and Repustar, have combined the use of AI and human fact-checking on social media, including for identification of election-related disinformation.[389] Logically worked on a government project involving alerting U.S. election officials to online disinformation intended to dissuade voting[390]; it also works with Facebook in the United Kingdom.[391]

The Partnership for AI has done a landscape review of platform interventions, posing fundamental questions about their use, including when each type should be used, who decides, what metrics and goals should inform auditing, and how their use can be made more transparent and trustworthy.[392] Several research reviews and studies on such interventions identified areas for further study and greater transparency, having concluded that we do not yet know what

---

[383] *See, e.g.*, https://debunk.eu/; https://fullfact.org/about/automated/.

[384] *See, e.g.*, Barrett, *Who Moderates the Social Media Giants?*, *supra* note 281 at 5.

[385] *See* Fatemeh Torabi Asr and Maite Taboada, *Big Data and quality data for fake news and misinformation detection*, Big Data and Society (Jan-Jun. 2019) (advocating for higher quality datasets), https://journals.sagepub.com/doi/10.1177/2053951719843310.

[386] *See* Ashkan Kazemi, et al., *Claim Matching Beyond English to Scale Global Fact-Checking* (Jun. 2021), https://arxiv.org/pdf/2106.00853.pdf.

[387] *See* https://maldita.es/uploads/public/docs/disinformation_on_whatsapp_ff.pdf; https://eu.boell.org/en/2021/10/04/inside-your-pocket-grave-threat-disinformation-private-messenger-apps. Similarly, although not involving such sophisticated technology, the encrypted messaging app Line (a popular communication platform in Asia) has partnered for several years with local fact-checking organizations and allows users to report suspicious messages and receive real-time answers about whether they're true, thus allowing for tracking of harmful content without breaking encryption. *See* Andrew Deck and Vittoria Elliott, *How Line is fighting disinformation without sacrificing privacy,* Rest of World (Mar. 7, 2021), https://restofworld.org/2021/how-line-is-fighting-disinformation-without-sacrificing-privacy/.

[388] *See* Jennifer Allen, et al., Scaling up fact-checking using the wisdom of crowds, ScienceAdvances 7:36 (Sep. 1, 2021), https://www.science.org/doi/10.1126/sciadv.abf4393; William Godel, et al., *Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking*, J. of Online Trust and Safety (Oct. 2021), https://tsjournal.org/index.php/jots/article/view/15/6.

[389] *See* https://www.logically.ai/about; https://repustar.com/events-resources. *See also* United Nations Development Programme, *In Honduras, iVerify partners with local university to support national elections* (Feb. 10, 2022), https://digital.undp.org/content/digital/en/home/stories/in-honduras--iverify-partners-with-local-university-to-support-n.html.

[390] *See* Rachael Levy, *Homeland Security Considers Outside Firms to Analyze Social Media After Jan. 6 Failure*, Wall St. J. (Aug. 15, 2021), https://www.wsj.com/articles/homeland-security-considers-outside-firms-to-analyze-social-media-after-jan-6-failure-11629025200.

[391] *See* https://www.logically.ai/press/logically-announces-uk-fact-checking-partnership-with-facebook.

[392] *See* Emily Saltz and Claire Leibowicz*, Fact-Checks, Info Hubs, and Shadow-Bans: A Landscape Review of Misinformation Interventions*, Partnership on AI (Jun. 14, 2021), https://www.partnershiponai.org/intervention-inventory/.

measures work, or work best, in what circumstances.[393] One study explored how multiple types of intervention may work better than one in isolation.[394] Further, University of Washington Professor Kate Starbird has explained that another challenge to the efficacy of a given intervention on one platform, such as the downranking of a YouTube video, is that the reach of that content is often driven by dynamics and engagement occurring on *other* platforms.[395] A similar challenge is that labeling or blocking content on one platform may result in that content proliferating elsewhere, as New York University researchers found with respect to certain election disinformation addressed by Twitter.[396]

The promises of, challenges with, and opacity regarding present use of platform interventions tend to lead naturally to questions of design and the larger social media ecosystem. For example, Professor Ellen P. Goodman has argued for policymakers to put virality disruptors and other types of content moderation into the context of user interface design.[397] Tel Aviv University Professor Niva Elkin-Koren has explored the possible use of "contesting algorithms" that would use adversarial design to mitigate some problems with AI-based content moderation, as well as a "separation of functions" for AI systems that would apply different oversight to systems collecting or labeling information than to those detecting or filtering content.[398] In its report on "information disorder," the Aspen Institute described alternative platform designs worthy of study, including two nonprofit examples that use AI algorithms: Pol.is, which works to bridge

---

[393] *See* Laura Courchesne, et al., *Review of social science research on the impact of countermeasures against influence operations*, Harvard Kennedy School Misinformation Rev. 2:5 (Sep. 2021), https://misinforeview hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/; Jon Bateman, et al., *Measuring the Efficacy of Influence Operations Countermeasures: Key Findings and Gaps From Empirical Research*, Carnegie Endowment for Int'l Peace (Sep. 2021), https://carnegieendowment.org/2021/09/21/measuring-efficacy-of-influence-operations-countermeasures-key-findings-and-gaps-from-empirical-research-pub-85389; William T. Adler and Dhanaraj Thakur, *A Lie Can Travel: Election Disinformation in the United States, Brazil, and France*, Center for Democracy and Technology, and Konrad Adenauer Stiftung (Dec. 2021), https://cdt.org/insights/cdt-and-kas-report-a-lie-can-travel-election-disinformation-in-the-united-states-brazil-and-france/.

[394] Joseph B. Bak-Coleman, et al., *Combining interventions to reduce the spread of viral misinformation* (May 23, 2021), https://osf.io/preprints/socarxiv/4jtvm.

[395] Kate Starbird, Twitter Post (Aug. 25, 2021), https://twitter.com/katestarbird/status/1430568134927208455?s=03. *See also* Mohan, *supra* note 234 (acknowledging YouTube's challenge to address this issue and suggesting use of interstitials before viewers can watch "a borderline embedded or linked video").

[396] *See* Zeve Sanderson, et al., *Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform*, Harvard Kennedy School Misinformation Rev. (Aug. 24, 2021), https://misinforeview hks.harvard.edu/article/twitter-flagged-donald-trumps-tweets-with-election-misinformation-they-continued-to-spread-both-on-and-off-the-platform/.

[397] *See* Ellen P. Goodman, *The Stakes of User Interface Design for Democracy* (Jul. 7, 2021), http://dx.doi.org/10.2139/ssrn.3882012. *See also* Sanderson, *supra* note 396; Will Oremus, *Facebook and YouTube's vaccine misinformation problem is simpler than it seems*, The Washington Post (Jul. 21, 2021), https://www.washingtonpost.com/technology/2021/07/21/facebook-youtube-vaccine-misinfo/.

[398] *See* Niva Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence*, Big Data & Society (Jul. 29, 2020), https://doi.org/10.1177/2053951720932296; Mayaan Perel and Niva Elkin-Koren, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 Lewis & Clark Law Rev. 857 (2020), http://dx.doi.org/10.2139/ssrn.3439261.

divided camps rather than maximizing engagement; and Local Voices Network, which helps community organizations facilitate constructive discussions.[399]

Uncovering networks and actors

Another way that platforms use AI tools to address online harms focuses not on the approach of identifying individual pieces of content but on finding the networks and actors behind them. With the aid of human intelligence about threats like criminal activity, TVEC, and election disinformation, sophisticated tools can map out patterns, signals, and behavioral indicators across many pieces of content and even across platforms.[400] For example, the Social Sifter project of North Carolina State University's Laboratory for Analytic Sciences involves using machine learning models to identify, track, and model foreign influence operations across social media platforms.[401] Google's Jigsaw and Facebook both make efforts to map and address coordinated inauthentic behavior (CIB).[402] Cross-platform mapping of certain communities, like those spreading online hate, is important because their members don't stay on one platform and move increasingly to smaller platforms featuring less content moderation.[403] However, tools that capture CIB may inadvertently ensnare minority groups or others who use protective methods to communicate on social media or via messaging apps about authoritarian regimes.[404]

---

[399] *See* Aspen Institute, *supra* note 8 at 47. *See also* https://pol.is/home; Audrey Tang, *A Strong Democracy Is a Digital Democracy*, The New York Times (Oct. 15, 2019), https://www.nytimes.com/2019/10/15/opinion/taiwan-digital-democracy html; https://cortico.ai/local-voices-network/; https://news mit.edu/2021/center-constructive-communication-0113.

[400] DARPA announced in May 2022 that it will explore the use of AI to map flows and identify patterns of influence operations across platforms. *See* https://sam.gov/opp/a28c282ea87f42568492247671580d0a/view. *See also* Elliott Waissbluth, et al., *Domain-Level Detection and Disruption of Disinformation* (May 6, 2022), https://arxiv.org/pdf/2205.03338v1.pdf; Samuel Woolley, *How Can We Stem the Tide of Digital Propaganda?*, CIGI (Jul. 5, 2021), https://www.cigionline.org/articles/how-can-we-stem-the-tide-of-digital-propaganda/; *Terrorism and Social Media: #IsBigTechDoingEnough?*, Sen. Comm. on Commerce, Science, and Transportation, 115th Cong. (2018) (testimony of Clint Watts), https://www.commerce.senate.gov/services/files/12847244-A89D-4A68-A6A5-CF9CB547E35B.

[401] *See* https://symposium ncsu-las net/influence html. *See also* Diogo Pacheco, et al., *Uncovering Coordinated Networks on Social Media: Methods and Case Studies*, Proc. AAAI Intl. Conf. on Web and Social Media (Apr. 7, 2021) (complementing measures to detect individual bot-driven or abusive accounts by looking at unexpectedly similar behavior of groups of actors, regardless of intent or automation), https://arxiv.org/abs/2001.05658.

[402] *See* Jigsaw, *Hate "Clusters" Spread Disinformation Across Social Media. Mapping Their Networks Could Disrupt Their Reach*, Medium (Jul. 28, 2021), https://medium.com/jigsaw/hate-clusters-spread-disinformation-across-social-media-995196515ca5; Nathaniel Gleicher, *Removing New Types of Harmful Networks*, Facebook (Sep. 16, 2021), https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/.

[403] *See, e.g.*, Alexandra T. Evans and Heather J. Williams, *How Extremism Operates Online* at 14, RAND Corp. (Apr. 2022), https://www rand.org/pubs/perspectives/PEA1458-2 html; Jigsaw, *supra* note 187; Candace Rondeaux, et al., *Parler and the Road to the Capitol Attack*, New America Future Frontlines (Jan. 5, 2022), https://www.newamerica.org/future-frontlines/reports/parler-and-the-road-to-the-capitol-attack/i-introduction. *See also* Drew Harwell and Will Oremus, *Only 22 saw the Buffalo shooting live. Millions have seen it since*, The Washington Post (May 16, 2022), https://www.washingtonpost.com/technology/2022/05/16/buffalo-shooting-live-stream/.

[404] *See* Zelly Martin, et al., *The K-Pop Fans Who Have Become Anti-Authoritarian Activists in Myanmar*, Slate (Oct. 21, 2021), https://slate.com/technology/2021/10/k-pop-fans-myanmar-activists.html.

Amplification of trustworthy content and counter-disinformation campaigns

An indirect way to address online harms is to increase user engagement with broadly trusted sources. Platforms and others can do so by generally amplifying such sources or specifically targeting those subject to harmful content or disinformation campaigns.[405] The State Department has expressed strong support of counter-disinformation measures, including debunking of false information.[406] The Surgeon General has advocated for technology firms to amplify information from trusted sources,[407] and both Facebook and YouTube have stated that they do so.[408] Dr. Erin Saltman of GIFCT has noted that providing counter-narratives or off-ramps to better information can be especially helpful for people who may be at high risk of succumbing to falsehoods but have not yet been converted to such views.[409] An example of this approach — one that relies in part on machine learning — is the Redirect Method, used by Moonshot and Google Jigsaw and discussed above in connection with TVEC.[410] Jigsaw and others are also studying the efficacy of prebunking, i.e., using technological tools to inoculate people against things like radicalization, extremism, and racism.[411] Two concerns with amplifying trustworthy content — or targeting and redirecting people to it — are who gets to decide what sources are authoritative, and to what extent will users believe them to be so.

F. User tools

Some platform design features and third-party services are or could be made available to help individuals avoid harmful or sensitive content on their own. Unlike the back-end interventions

---

[405] While the focus of such measures is often on platforms, others advocate for the primary role of civil society organizations in counter-disinformation measures. *See* Kevin Shieves, *How to Support a Globally Connected Counter-Disinformation Network*, War on the Rocks (Jan. 20, 2022) (noting that some of these groups use algorithms and other advanced tools and that generally they need appropriate data access, training, and funding), https://warontherocks.com/2022/01/how-to-support-a-globally-connected-counter-disinformation-network/.

[406] *See* https://www.state.gov/Disarming-Disinformation/.

[407] *See* Surgeon General, *supra* note 242 at 12.

[408] *See* https://about fb.com/news/2018/01/trusted-sources/; https://blog.youtube/inside-youtube/tackling-misinfo; https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

[409] Saltman, et al., *New Models for Deploying Counterspeech, supra* note 185.

[410] *See* https://moonshotteam.com/the-redirect-method/; https://jigsaw.google.com/the-current/white-supremacy/countermeasures/; Emily Dreyfuss, *Hacking Online Hate Means Talking to the Humans Behind It*, WIRED (Jun. 8, 2017), https://www.wired.com/2017/06/hacking-online-hate-means-talking-humans-behind/.

[411] *See, e.g.,* Ullrich K. H. Ecker, et al., *The psychological drivers of misinformation belief and its resistance to correction*, Nature Reviews Psychology 1: 13-29 (Jan. 12, 2022), https://www nature.com/articles/s44159-021-00006-y; NATO Strategic Communications Centre of Excellence, *Inoculation Theory and Misinformation* (Oct. 2021), https://stratcomcoe.org/pdfjs/?file=/publications/download/Inoculation-theory-and-Misinformation-FINAL-digital-ISBN-ebbe8.pdf?zoom=page-fit; Beth Goldberg, *Psychological Inoculation: New Techniques for Fighting Online Extremism*, Jigsaw (Jun. 24, 2021), https://medium.com/jigsaw/psychological-inoculation-new-techniques-for-fighting-online-extremism-b156e439af23; John Roozenbeek, et al., *Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures*, Harvard Kennedy School Misinformation Rev. (Feb. 3, 2020), https://misinforeview hks.harvard.edu/article/global-vaccination-badnews/.

described above, we refer here to tools that give users options for content control.[412] To the extent these tools involve identifying such content beyond techniques such as keyword searches or hash-matching, it is possible, if not likely, that AI would be working in the background.

In its open internet principles, Twitter advocated for prioritizing "human choice and control" over algorithms.[413] In 2021, its former CEO advocated for building a "marketplace" of social media algorithms,[414] and it announced that it would allow third-party developers to build programs atop Twitter to help, for example, with promoting healthy conversations.[415] The company has also rolled out and suggested ideas for new tools that would filter and limit potentially harmful replies and comments that users don't want to see.[416] One third-party app currently working on Twitter is Block Party, which attempts to filter harassing comments.[417] Instagram, too, has introduced features to help users hide or block potentially harmful content appearing in comments or direct messages.[418] Frustrated users of the Twitch gaming platform have created their own tools designed to limit harassment and hate in the user chat feature.[419]

Stanford University Professor Francis Fukuyama has proposed that an answer for harms attributable to social media platforms would be an open market in which users choose between independent filtering services, rather than rely on a platform's algorithmic determinations of what one will see.[420] A group of experts debated this "middleware" proposal in a set of articles in the Journal for Democracy and a follow-up conference.[421] Some expressed cautious optimism,

---

[412] *See, e.g.*, CFDD, *supra* note 224 at 18 (advocating for introduction of such tools); Aspen Institute, *supra* note 8 at 66-67 (same).

[413] *See* Twitter, *Protecting the Open Internet*, *supra* note 377 at 3. *See also* Kate Conger, *Twitter Wants to Reinvent Itself, by Merging the Old with the New*, The New York Times (Mar. 2, 2022), https://www.nytimes.com/2022/03/02/technology/twitter-platform-rethink.html; Field, supra note 319 (comments of Twitter's Rumman Chowdhury supporting "algorithmic choice" for consumers despite implementation challenges).

[414] *See* Jacob Kastrenakes, *Twitter's Jack Dorsey wants to build an app store for social media algorithms*, The Verge (Feb. 9, 2021), https://www.theverge.com/2021/2/9/22275441/jack-dorsey-decentralized-app-store-algorithms.

[415] *See* https://blog.twitter.com/developer/en_us/topics/tools/2021/build-whats-next-with-the-new-twitter-developer-platform.

[416] *See* https://blog.twitter.com/en_us/topics/product/2021/introducing-safety-mode; Ian Carlos Campbell, *Twitter seeking input as it explores Filter and Limit controls on tweets*, WIRED (Sep. 24, 2021), https://www.theverge.com/2021/9/24/22692264/twitter-filter-limit-tweet-replies-automatic.

[417] *See* https://www.blockpartyapp.com/.

[418] *See, e.g.*, https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse; https://about.instagram.com/blog/announcements/introducing-sensitive-content-control.

[419] *See* Ash Parrish, *How to stop a hate raid*, The Verge (Aug. 20, 2021), https://www.theverge.com/22633874/how-to-stop-a-hate-raid-twitch-safety-tools.

[420] *See* Francis Fukuyama, *Making the Internet Safe for Democracy*, Journal of Democracy 32:2 (Apr. 2021), https://muse.jhu.edu/article/787834. *See also* Future of Tech Commission, *supra* note 297 at 31.

[421] *See* Journal of Democracy 32:3 (Jul. 2021), https://muse.jhu.edu/issue/44978; Richard Reisman, *Progress Toward Re-Architecting Social Media to Serve Society*, Tech Policy Press (Dec. 1, 2021), https://techpolicy.press/progress-toward-re-architecting-social-media-to-serve-society/.

pointing to similar ideas in the past,[422] but others see no viable business model, technological impediments, or problems with speech, privacy, and competition.[423] Although middleware may be mostly an idea only, a relatively new third-party service that might qualify is Preamble, an AI-based option for Twitter that adjusts rankings in accord with users' selections of "values providers."[424]

Tools that give users more control and information, along with amplifying trustworthy content and engaging in debunking and prebunking efforts, are all closely aligned with the idea of promoting digital literacy. An important element of a whole-of-society approach to countering online harms, digital literacy is the subject of many projects, policy proposals, and research.[425] Two recent studies indicate that improving digital literacy skills shows promise against different kinds of online disinformation.[426] Further, reports commissioned by DHS stress that building public resilience to such content may ultimately be more effective than focusing on technological solutions.[427] One application, supported by both DHS and the State Department, is a free browser game, Harmony Square, which draws on "inoculation theory" to get people to appreciate techniques used in online misinformation surrounding elections and thus make them

---

[422] *See, e.g.*, Mike Masnick, *Protocols, Not Platforms: A Technological Approach to Free Speech*, Knight First Amendment Institute (Aug. 21, 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech; Stephen Wolfram, *Testifying at the Senate about A.I.-Selected Content on the Internet* (Jun. 25, 2019), https://writings.stephenwolfram.com/2019/06/testifying-at-the-senate-about-a-i-selected-content-on-the-internet/.

[423] *Id.*

[424] *See* https://www.preamble.ai/about-us.

[425] *See, e.g.*, Future of Tech Commission, *supra* note 297 at 14, 20, 23; Royal Society, *supra* note 359 at 21, 84; Aspen Institute, *supra* note 8 at 64-68; Kristin M. Lord and Katya Vogt, *Strengthen Media Literacy to Win the Fight Against Misinformation*, Stanford Soc. Innov. Rev. (Mar. 18, 2021), https://ssir.org/articles/entry/strengthen_media_literacy_to_win_the_fight_against_misinformation#; P.W. Singer and Michael McConnell, *Want to Stop the Next Crisis? Teaching Cyber Citizenship Must Become a National Priority*, TIME (Jan. 21, 2021), https://time.com/5932134/cyber-citizenship-national-priority/. Of course, educational efforts have their limits, as even the most digitally literate consumers cannot reasonably avoid all types of online harms. *See* Monica Bulger and Patrick Davison, *The Promises, Challenges and Futures of Media Literacy*, J. of Media Literacy Education 10(1) (2018), https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1365&context=jmle.

[426] *See* Bertie Vidgen, et al., *Understanding vulnerability to online misinformation* 5-6, The Alan Turing Institute (Mar. 2021), https://www.turing.ac.uk/sites/default/files/2021-02/misinformation_report_final1_0.pdf; Andrew M. Guess, et al., *A digital media literacy intervention increases discernment between mainstream and false news in the United States and India*, PNAS 117(27): 15536-45 (Jul. 7, 2020), www.pnas.org/cgi/doi/10.1073/pnas.1920498117.

[427] *See* DHS Analytic Exchange Program, *Combatting Targeted Disinformation Campaigns, Part Two* at 38-43; DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 31. As Lawrence Krauss theorized, the internet will continue to "propagate out of control" no matter what businesses and governments do, so "becoming your own filter will become the challenge of the future." Lawrence Krauss, *Lo and Behold*, directed by Werner Herzog. Chicago: Saville Productions, 2016.

better able to resist it.[428] Some countries, including the United Kingdom, are incorporating digital literacy as part of concerted national strategies.[429]

G. Availability and scalability

We have noted already some of the problems with the fact that only a few large technology companies are responsible for most of the AI tools within the scope of this report. For any such tool that is effective and fair, another problem with this concentration is that others who may need the tool, like smaller platforms or investigative journalists, won't necessarily have access to it or the resources to create their own.[430] Twitter even discusses this problem in its open internet principles, advocating for more accessibility and regretting that such technology remains in "proprietary silos" and that this fact perpetuates the domination of a few companies.[431]

Greater access to these tools does carry risk. For example, while sharing an algorithm may not involve exposure of personal information, sharing the dataset used to create an AI model could implicate privacy concerns. Such concerns may be more acute when the sharing is with other commercial actors as opposed to vetted researchers or certified auditors. Sharing technology and information also risks cross-site censorship.[432] Further, the more widely a detection or mitigation

---

[428] *See* Jon Roozenbeek and Sander Van Der Linden, *Breaking Harmony Square: A game that "inoculates" against political misinformation*, Harvard Kennedy School Misinformation Rev. (Nov. 6, 2020), https://misinforeview hks.harvard.edu/article/breaking-harmony-square-a-game-that-inoculates-against-political-misinformation/. *See also* Nicholas Micallef, et al., *Fakey: A Game Intervention to Improve News Literacy on Social Media,* Proc. ACM Hum.-Comput. Interact., Vol. 5, No. CSCW1 (Apr. 2021), https://dl.acm.org/doi/10.1145/3449080.

[429] *See* https://www.gov.uk/government/publications/online-media-literacy-strategy; Amy Yee, *The country inoculating against disinformation*, BBC Future (Jan. 30, 2022) (showing the positive effects of such efforts in Estonia), https://www.bbc.com/future/article/20220128-the-country-inoculating-against-disinformation.

[430] *See, e.g.*, UK Dept. for Digital, Culture, Media & Sport, *Understanding how platforms with video-sharing capabilities protect users from harmful content online* (Aug. 2021), https://www.gov.uk/government/publications/understanding-how-platforms-with-video-sharing-capabilities-protect-users-from-harmful-content-online; Royal Society, *supra* note 359 at 18, 82. These needs are often discussed in the TVEC and deepfake contexts. *See, e.g.,* also DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 31; Tech Against Terrorism, *GIFCT Technical Approaches Working Group Gap Analysis and Recommendations* at 24-25; Jacob Berntsson and Maygane Janin, *Online Regulation of Terrorist and Harmful Content*, Lawfare (Oct. 14, 2021), https://www.lawfareblog.com/online-regulation-terrorist-and-harmful-content; OECD, *Transparency Reporting on Terrorist and Violent Content Online*, *supra* note 190 at 12; EPRS, *Tackling deepfakes in European policy*, *supra* note 43 at 59.

[431] *See* Twitter, *Protecting the Open Internet*, *supra* note 377 at 8.

[432] *See* Emma Llansó, *Content Moderation Knowledge Sharing Shouldn't Be a Backdoor to Cross-Platform Censorship*, TechDirt (Aug. 21, 2020), https://www.techdirt.com/articles/20200820/08564545152/content-moderation-knowledge-sharing-shouldnt-be-backdoor-to-cross-platform-censorship.%E2%80%A6.

tool is shared, the easier it will be for bad actors to exploit, meaning that dissemination should be controlled carefully.[433]

H. Content authenticity and provenance

Given the many difficulties with using AI or other automated means to detect harmful content, it makes sense to focus on the flip side: authentication. While authentication tools do not necessarily help with every harm listed by Congress, they can be widely used to help determine the true source of content and whether text, images, audio, or video are deepfakes (see above) or have been otherwise manipulated. Indeed, multiple federal government reports state that these tools are key for challenging foreign disinformation and deepfakes.[434] Experts from the State Department and elsewhere have pointed to blockchain technology as a means of determining content authenticity.[435] Authentication could also help counteract the Liar's Dividend, a problem discussed above, in that it would be harder for public figures to claim falsely that audio or video content is fake if one could point to technological markers that it is real and unaltered.

A major collaborative effort to advance authentication tools is the Coalition for Content Provenance for Authenticity (C2PA), formed in early 2021 by merging two other coordinated efforts, the Content Authenticity Initiative (led by Adobe) and Project Origin (led by Microsoft and the BBC). The goal of this coalition is to create an "open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media."[436] In January 2022, it released technical specifications and guidance documents.[437]

Of course, proving that content has not been altered and comes from its claimed origin does not prove the truth of the content itself. Further, and just like detection technology, these tools are fallible, and it would be problematic if people were either too distrustful of content that had no authenticity markers or too trusting of content that did. For example, authentication does not help

---

[433] This issue is discussed above in the part of Section I on deepfakes. *See also* Sam Gregory, et al., *Governing Access to Synthetic Media Detection Technology*, Tech Policy Press (Sep. 7, 2021), https://techpolicy.press/governing-access-to-synthetic-media-detection-technology/; EPRS, *Tackling deepfakes in European policy*, *supra* note 43 at 59 .

[434] *See* NSCAI, *Final Report*, *supra* note 3 at 48; DHS, *Increasing Threat of Deepfake Identities*, *supra* note 43 at 31; EPRS, *Tackling deepfakes in European policy*, *supra* note 43 at 20, 65. *See also* Jaiman, *supra* note 62; Engler, *supra* note 62.

[435] *See* J.D. Maddox, et al., *Toward a More Ethical Approach to Countering Disinformation Online*, Public Diplomacy 23(12) (Jul. 1, 2020), https://static1.squarespace.com/static/5be3439285ede1f05a46dafe/t/5efd72972af517215e330cdd/1593668272484/ETHICS+IN+DIPLOMACY+Final.pdf. *See also* Kathryn Harrison and Amelia Leopold, *How Blockchain Can Help Combat Disinformation*, Harvard Bus. Rev. (Jul. 19, 2021), https://hbr.org/2021/07/how-blockchain-can-help-combat-disinformation; Haya R. Hasan and Khaled Salah, *Combating Deepfake Videos Using Blockchain and Smart Contracts*, IEEE Access 7:41596 (Feb. 25, 2019), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8668407. The News Provenance Project is also exploring the use of blockchain as a way to store contextual information about news photos. *See* https://www newsprovenanceproject.com/a-solution.

[436] *See* https://c2pa.org/.

[437] *See* https://contentauthenticity.org/blog/milestones-in-digital-content-provenance-specification-open-source-projects.

with "shallowfakes" — when malicious actors upload real and unaltered media but change the context and claim it depicts different people at different places or times.[438] It is also possible that people could abuse these tools, extracting data from them and using them for surveillance.[439]

As authentication tools advance, and especially as they scale, it is important to ensure that they enhance trust and freedom of expression, not harm it. Sam Gregory, Program Director of WITNESS, points out that human rights activists, lawyers, media outlets, and journalists "often depend for their lives on the integrity and veracity of images they share from conflict zones, marginalized communities and other places threatened by human rights violations."[440] Sometimes, however, whether to protect themselves or their subjects, they may need to use pseudonyms, blur faces, or obscure locations.[441] We would not want authentication systems to block the resulting videos or for viewers to ignore them because they lack certain markers.

I. Legislation

Legislative efforts around the world may reflect that the only effective ways to deal with online harm are laws that change the business models or incentives allowing harmful content to proliferate. Under debate in Congress are, among other things, proposals involving Section 230 of the Communications Decency Act, data privacy, and competition. Some of these proposals give the FTC new responsibilities. Nonetheless, Congress did not seek recommendations on how to deal with online harm generally, so these proposals are beyond the bounds of this report.

The Congressional request is narrower. It asks the FTC to recommend laws that would "advance the adoption and use of artificial intelligence to address" the listed online harms. In fact, platforms and others already use AI tools to attempt to address most of those harms, but these tools are often neither robust nor fair enough to mandate or encourage their use. We look instead to the development of legal frameworks that would help ensure that such use of AI does not itself cause harm.[442]

---

[438] *See* Bobbie Johnson, *Deepfakes are solvable—but don't forget that "shallowfakes" are already pervasive*, MIT Tech. Rev. (Mar. 25, 2019), https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/.

[439] *See* Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all*, WITNESS Blog (May 2020), https://blog.witness.org/2020/05/authenticity-infrastructure/.

[440] *Id.*

[441] *See id.*

[442] While some existing laws may provide guardrails for some harms caused by some AI tools discussed herein, those laws are insufficient. *See, e.g.*, Slaughter, *supra* note 13 at 48; Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315 (2020), https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf; Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, Harv. J. L. & Tech. 31:2 (2018), https://jolt.law harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf.

Congress should generally steer clear of laws that require, assume the use of, or pressure companies to deploy AI tools to detect harmful content.[443] As discussed above, such tools are rudimentary and can result in bias and discrimination. Further, laws that push platforms to rapidly remove certain types of harmful content may not survive First Amendment scrutiny in any event, as they would tend both to result in the overblocking of lawful speech and impinge on platform discretion to determine editorial policies,[444] concerns that do not prevent such laws in countries without that constitutional restriction.[445] We note also that asking platforms and other private actors to make quick decisions about the illegality of content is in jarring contrast to the amount of time and deliberation that courts and agencies use to make similar decisions.[446] On the other hand, some of these concerns are less present for certain categories like CSAM, fraud, and illegal product sales, as to which quick takedown requirements may be desirable and less controversial.

For any law that does address AI use and online harm, three critical considerations are definitions, coverage, and offline effects. First, difficulties arise in defining both technological terms and the harms to be addressed. As explained above, definitions of terms like AI and algorithm are highly problematic because of their ambiguity and breadth. Congress can employ better terminology as applicable, like Rashida Richardson's specific proposal to use "automated

---

[443] *See, e.g.*, Shenkman, *supra* note 224 at 36; Gorwa, *supra* note 224 at 2-3; Bloch-Wehba, *supra* note 240 at 74-87; Duarte, *supra* note 258 at 14-15.

[444] *See, e.g.*, Future of Tech Commission, *supra* note 297 at 19, 21; Daphne Keller, *Amplification and Its Discontents*, Knight First Amendment. Institute (Jun. 8, 2021), https://knightcolumbia.org/content/amplification-and-its-discontents; Emma Llansó, et al., *Artificial Intelligence, Content Moderation, and Freedom of Expression*, Transatlantic Working Group (Feb. 26, 2020) (arguing that governments should "resist simplistic narratives about all-powerful algorithms or AI as being the sole cause of, or solution to, the spread of harmful content online"), https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf; Singh, *Everything in Moderation*, *supra* note 224 at 33; Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money*, Hoover Institution Aegis Paper Series (Jun. 3, 2018), https://www hoover.org/research/internet-platforms-observations-speech-danger-and-money.

[445] Several foreign laws and proposals effectively mandate algorithmic detection methods and quick takedowns for certain types of content. *See, e.g.*, Daphne Keller, *Five Big Problems with Canada's Proposed Regulatory Framework for "Harmful Online Content,"* Tech Policy Press (Aug. 31, 2021), https://techpolicy.press/five-big-problems-with-canadas-proposed-regulatory-framework-for-harmful-online-content/?s=03; evelyn douek, *Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech*, 94 Australian L. J. 41 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3443220; Bloch-Wehba, *supra* note 240 at 83-85; Elkin-Koren, *Contesting Algorithms*, *supra* note 240 at 2-3; https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/; https://www.liberties.eu/en/stories/terrorist-content-regulation-open-letter-to-meps/43410; Joris van Hoboken, *The Proposed EU Terrorism Content Regulation*, Transatlantic Working Group (May 3, 2019), https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/EU_Terrorism_Regulation_TWG_van_Hoboken_May_2019.pdf.

[446] *See, e.g.*, Jacob Mchangama, *Rushing to Judgment: Examining Government Mandated Content Moderation*, Lawfare (Jan. 26, 2021), https://www.lawfareblog.com/rushing-judgment-examining-government-mandated-content-moderation.

decision systems," or it can attempt to avoid the issue by focusing on outcomes and impacts.[447] Defining any given harm can also be problematic, however, such as when it has no existing definition in federal law, when a legal definition exists but does not translate to the online context, or when the harm itself is amorphous and subject to different meanings depending on the context. Second, the law's scope is also crucial. What parts of the tech stack are covered? If limited to social media companies, should it distinguish between such companies based on size, and how should size be measured?[448] Any law that effectively mandates automated tools could serve to benefit the few platforms that have the financial and technological means of compliance, increasing the barriers that new entrants would need to overcome.[449] Congress should also consider generational changes in what people use to communicate online and avoid covering only services that a particular generation is using right now and that might diminish in popularity over time.[450] Third, online harms have offline dimensions, not only because harmful events in the physical world serve as the impetus for online content but also because — as noted above in the discussion of hate speech — online content can have serious offline consequences. Legislators should thus avoid treating online harm in isolation.

As previewed above, we believe any initial legislative focus should prioritize the transparency and accountability of platforms and others that build and use automated systems to address online harms. Again, while this approach may not itself solve or reduce those harms, it would allow policymakers, researchers, and the public to understand the use and impact of those tools and provide evidence for what measures should follow.[451] While some platforms provide helpful information, at this point it seems clear that only legislation will allow us to crack open the black boxes of content moderation and the nesting black boxes of AI tools powering it.

The view that we need laws relating to algorithmic transparency and accountability — particularly for social media platforms and other technology companies — typically includes calls for: (1) public disclosure of information, including policies and data on the use and impact of AI systems; (2) researcher access to additional information; (3) protections for whistleblowers, auditors, researchers, and journalists; (4) requirements for audits and impact assessments; and (5) systems for flagging violative content and for notice, appeal, and redress for

---

[447] *See* Richardson, *Defining and Demystifying ADS*, *supra* note 7; Lum and Chowdhury, *What is an algorithm, supra* note 6; Spandana Singh, *Regulating Platform Algorithms*, New America (Dec. 1, 2021) (comparing current EU and US approaches to regulating platform algorithms), https://www.newamerica.org/oti/briefs/regulating-platform-algorithms/.

[448] *See* Eric Goldman and Jess Miers, *Regulating Internet Services by Size*, CPI Antitrust Chronicle, Santa Clara Univ. Legal Studies Research Paper (May 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3863015.

[449] *See, e.g.*, Bloch-Wehba, *supra* note 240 at 47, 87.

[450] *See* Mark MacCarthy, *Coming Soon to a Podcast, an App Store and a Metaverse Near You….Content Moderation Rules*, Forbes (Feb. 3, 2022), https://www forbes.com/sites/washingtonbytes/2022/02/03/coming-soon-to-a-podcast-an-app-store-and-a-metaverse-near-you-content-moderation-rules/.

[451] *See* evelyn douek, *Content Moderation as Administration*, 136 Harv. L. Rev. __ (forthcoming 2022) (arguing that pursuing regulation focusing on accountability is a first, pragmatic step towards any substantive reform), https://ssrn.com/abstract=4005326.

individuals affected by content removal or non-removal decisions.[452] We agree that each of these elements would be valuable components of any relevant legislation, but we would urge Congress to carefully consider the privacy and security risks that accompany enhanced access to data.[453] Two recent, noteworthy proposals for legislation are from Stanford University Professor Nathan Persily and Deborah Raji and concern mandated but controlled data access for researchers and

[452] *See, e.g.*, Slaughter, *supra* note 13 at 48-51; Future of Tech Commission, *supra* note 297 at 13, 22; Competition and Markets Authority, *supra* note 74 at 49-50; CFDD, *supra* note 224 at 26-29; Brennan Center for Justice, *supra* note 257 at 18-23; Paul M. Barrett, et al., *Fueling the Fire: How Social Media Intensifies U.S. Political Polarization —And What Can Be Done About It*, NYU Stern Center for Business and Human Rights at 23-24 (Sep. 2021), https://bhr.stern.nyu.edu/polarization-report-page; Singh and Doty, *Cracking Open the Black Box*, *supra* note 310 at 33-35; Llansó, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, *supra* note 444 at 25; Bloch-Wehba, *supra* note 240 at 87-94; Daphne Keller, *Some Humility about Transparency*, The Center for Internet and Society (Mar. 19, 2021) (referring to effects on midsized or small platforms), http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency. Amba Kak and Rashida Richardson, *Artificial Intelligence Policies Must Focus on Impact and Accountability* (May 1, 2020), https://www.cigionline.org/articles/artificial-intelligence-policies-must-focus-impact-and-accountability/; evelyn douek, *Facebook's White Paper on the Future of Online Content Regulation: Hard Questions for Lawmakers*, Lawfare (Feb. 18, 2020), https://www.lawfareblog.com/facebooks-white-paper-future-online-content-regulation-hard-questions-lawmakers; Mark MacCarthy, *How online platform transparency can improve content moderation and algorithmic performance*, Brookings TechTank (Feb. 17, 2021), https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/; Mark McCarthy, *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*, Transatlantic Working Group (Feb. 12, 2020), https://www.ivir nl/publicaties/download/Transparency_MacCarthy_Feb_2020.pdf; *Task Force on Artificial Intelligence*, *supra* note 316 (testimony of Meredith Broussard, Miriam Vogel, and Aaron Cooper); *H. Comm. on Science, Space, and Technology* (testimony of Meredith Whittaker), *supra* note 252 at 12-15; Twitter, *Protecting the Open Internet* , *supra* note 377 at 5-10 (regulation should focus on "system-wide processes," noting that problems stem from "platform design choices that are dictated by business models," and arguing that transparency and accountability methods would let us know what kind of laws and interventions would actually be effective). *But see* Eric Goldman, *The Constitutionality of Mandating Editorial Transparency,* 73 Hastings L. J. __ (2022) (forthcoming) (arguing that laws mandating editorial transparency may violate the First Amendment and that legal reform should focus on certified and independent audits, researcher scraping, and increased digital citizenship education), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005647.

[453] Several of these elements are key provisions of both the European Union's proposed Digital Services Act and the United Kingdom's Online Safety Bill. *See* https://www.europarl.europa.eu/news/en/press-room/20211210IPR19209/digital-services-act-safer-online-space-for-users-stricter-rules-for-platforms; https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/12902 htm?s=03. *See also* Alex Engler, *Platform data access is a lynchpin of the EU's Digital Services Act*, Brookings TechTank (Jan. 15, 2021), https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/.

auditors, respectively.[454] Definitions and standard-setting are also important in this area and should not be limited to technical disciplines and concepts.[455]

We are aware of, and are encouraged by, Congressional bills that move in these directions, and we would be happy to engage with Congress on any such bills that proceed. Indeed, some of these bills provide roles for the FTC, as to which we express hope that Congress will consider addressing relevant agency resource needs in conjunction with adding any new responsibilities.

V. CONCLUSION

"Platforms dream of electric shepherds," says Tarleton Gillespie, expressing skepticism that automation can replace humans in addressing harmful online content.[456] Legislators and regulators with similar dreams should remain skeptical as well. Dealing effectively with online harms requires substantial changes in business models and practices, along with cultural shifts in how people use or abuse online services. These changes involve significant time and effort across society and can include, among other things, technological innovation, transparent and accountable use of that technology, meaningful human oversight, global collaboration, digital literacy, and appropriate regulation. AI is no magical shortcut.

---

[454] *See Social Media Platforms and the Amplification of Domestic Extremism and Other Harmful Content*, S. Comm. on Homeland Security and Governmental Affairs, 117th Cong. (2021) (testimony of Nathan Persily), https://www.hsgac.senate.gov/imo/media/doc/Testimony-Persily-2021-10-28.pdf; Deborah Raji, *Third-Party Auditor Access for AI Accountability*, in Policy and AI: Four Radical Proposals for a Better Society, Stanford HAI (Nov. 2021) (also suggesting certifications, auditor oversight board, and national incident reporting system), video available at https://hai.stanford.edu/news/radical-proposal-third-party-auditor-access-ai-accountability.

[455] *See, e.g.*, Brandie Nonnecke and Philip Dawson, *Human Rights Implications of Algorithmic Impact Assessments*, Carr Center for Human Rights Policy, Harvard Kennedy School (Fall 2021), https://carrcenter hks.harvard.edu/publications/human-rights-implications-algorithmic-impact-assessments-priority-considerations. Dr. Mona Sloane noted that the recent focus on audits means we need to define the term and specify its scope, or else we will "see lots of audit-washing in industry, lots of random audit-labeling in research, and no real change." Mona Sloane, Twitter Post (Dec. 22, 2021), https://twitter.com/mona_sloane/status/1473559128253546501?t=m1tLfxFzMnF373shIKTyKg&s=03.

[456] Gillespie, Custodians of the Internet, *supra* note 225 at 107-08.

Acknowledgements

The drafter of this report is Michael Atleson of the Bureau of Consumer Protection. Additional acknowledgement goes to Sarah Myers West, Amba Kak, and Olivier Sylvain, all of whom are advisors to the Chair, as well as to Elisa Jillson, Robin Wetherill, Ellen Connelly, Daniele Apanaviciute, Tawana Davis, and Serena Viswanathan, all of whom are from the Bureau of Consumer Protection.