CENTER FOR DEMOCRACY & TECHNOLOGY

June 12, 2023

To: National Telecommunications and Information Administration
Department of Commerce
1401 Constitution Avenue, NW, Room 4725
Washington, DC 20230

Re: AI Accountability Policy Request for Comment, NTIA-2023-005

Authors: Will Adler, Samir Jain, Elizabeth Laird, Emma Llansó, Nathalie Maréchal, Eric Null, Matt Scherer, Ridhi Shetty, Hannah Quay-de la Vallee. With thanks to: Ariana Aboulafia, Jared Katzman, Aidan Kierans, Ophélie Stockhem.

Table of Contents

Introduction

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to the National Telecommunications and Information Administration's (NTIA) request for comments regarding artificial intelligence ("AI") system accountability measures and policies. CDT is a nonprofit 501(c)(3) organization fighting to advance civil rights and civil liberties in the digital age. CDT's focus includes the impact of data- and algorithm-driven discrimination, as well as accountability for the entities involved in developing and deploying such systems. In addition to these comments, we invite the NTIA to refer to our recent agency comments and publications on worker surveillance,[1] tenant screening,[2] the relationship between trade negotiations and AI accountability,[3] large language models' performance in languages other

[1] CDT intends to file comments in the worker surveillance proceeding at the Office of Science and Technology and Policy on or before June 29.
[2] Ridhi Shetty, *CDT Comments to Federal Agencies Highlight Risks of Data Used in Tenant Screening*, Center for Democracy & Technology (June 2, 2023), https://cdt.org/insights/cdt-comments-to-federal-agencies-highlight-risks-of-data-used-in-tenant-screening/.
[3] Eric Null, *CDT and Civil Society Groups Warn Trade Talks Could Hurt Fight Against Discriminatory Algorithms*, Center for Democracy & Technology (May 25, 2023), https://cdt.org/insights/cdt-and-civil-society-groups-warn-trade-talks-could-hurt-fight-against-discriminatory-algorithms/.

than English,[4] digital identity,[5] risk assessments for automated decision-making,[6] public housing,[7] and the use of generative AI in education,[8] all of which pertain to aspects of AI accountability in specific sectors. The recent testimonies of CDT President and CEO Alexandra Reeve Givens before the Senate Committee on Homeland Security & Governmental Affairs,[9] the House Committee on Energy and Commerce,[10] and the Senate Committee on the Judiciary,[11] also address relevant topics.

The comments below begin with an illustrative (though not exhaustive) overview of AI-associated harms. A key goal for accountability should be to minimize these harms and to provide redress when they occur. Because "artificial intelligence" is such a capacious and flexible phrase,[12] we specifically address three types of algorithmic systems in turn: automated

---

[4] *See generally,* Gabriel Nicholas & Aliya Bhatia, *Lost in Translation: Large Language Models in Non-English Content Analysis*, Center for Democracy & Technology (May 23, 2023), https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.

[5] Hannah Quay-de la Valle, *CDT Joined by AJL, Researchers in NIST Comments Furthering Equity and Privacy in Digital Identity Guidelines*, Center for Democracy & Technology (Apr. 14, 2023), https://cdt.org/insights/cdt-joined-by-ajl-researchers-in-nist-comments-furthering-equity-and-privacy-in-digital-identity-guidelines/.

[6] Ridhi Shetty & Matt Scherer, *CDT Comments to California Privacy Protection Agency on Automated Decision-making and Risk Assessments*, Center for Democracy & Technology (March 28, 2023), https://cdt.org/insights/cdt-comments-to-california-privacy-protection-agency-on-automated-decision-making-and-risk-assessments/.

[7] *See* Michael Yang, *The Promise and Peril of Data & Technology Use by Public Housing Agencies*, Center for Democracy & Technology (March 2023), https://cdt.org/wp-content/uploads/2023/03/2023-03-17-Civic-Tech-Technology-in-Public-Housing-Agencies-final.pdf.

[8] Hannah Quay-de la Vallee, *Generative AI Systems in Education - Uses and Misuses*, Center for Democracy & Technology (Mar. 15, 2023) https://cdt.org/insights/generative-ai-systems-in-education-uses-and-misuses/.

[9] Alexandra Reeve Givens, *CDT CEO Alexandra Givens Testimony Before Senate Committee on Homeland Security & Governmental Affairs on "Artificial Intelligence: Risks and Opportunities,"* Center for Democracy & Technology (Mar. 8, 2023), https://cdt.org/insights/cdt-ceo-alexandra-givens-testimony-before-senate-committee-on-homeland-security-governmental-affairs-on-artificial-intelligence-risks-and-opportunities/.

[10] Alexandra Reeve Givens, *CDT CEO Alexandra Givens Testimony Before House Energy & Commerce Hearing on "Promoting U.S. Innovation and Individual Liberty Through a National Standard for Data Privacy*," Center for Democracy & Technology (Mar. 1, 2023), https://cdt.org/insights/cdt-ceo-alexandra-givens-testimony-before-house-energy-commerce-hearing-on-promoting-u-s-innovation-and-individual-liberty-through-a-national-standard-for-data-privacy/.

[11] Alexandra Reeve Givens, *CDT CEO Alexandra Givens Testimony Before Senate Judiciary on Artificial Intelligence and Human Rights*, Center for Democracy & Technology (June 13, 2023), https://cdt.org/insights/cdt-ceo-alexandra-givens-testimony-before-senate-judiciary-on-artificial-intelligence-and-human-rights/.

[12] "[W]hat exactly is 'artificial intelligence' anyway? It's an ambiguous term with many possible definitions. It often refers to a variety of technological tools and techniques that use computation to perform tasks such as predictions,

decision-making tools; systems for content analysis, moderation, and recommendation; and generative AI including large language models (LLMs). We discuss how the algorithmic systems in question are used in the public and private sectors, and the ensuing harms to individuals and groups of people. The seriousness of the potential harms makes clear that accountability for them cannot be left simply to industry, but must involve all key stakeholders, from the government and regulators to civil society and independent experts to communities and individuals harmed or otherwise affected by the use of AI.

Next, we present some high-level considerations for AI accountability policy. We discuss four major components of the AI accountability toolbox: transparency, explainability, and interpretability; audits and assessments; laws and liability; and government procurement reform.

- Transparency is the foundation of accountability. That starts with the disclosure that AI is being used in, for example, a decision about benefits. Of course, merely knowing the role of AI is not sufficient. Transparency also requires the disclosure of information such as how an AI system was trained, how it arrives at decisions, and an explanation of the decision or output in a particular case.

- Audits and assessments are widely understood as fundamental to accountability but important questions remain unanswered. These include (a) how to ensure auditors have sufficient independence, expertise, and resources; and (b) how to develop the standards to be used by auditors, recognizing that they will embody value judgments that should be made only after input from all affected stakeholders.

- In many cases, existing laws such as civil rights statutes provide basic rules that continue to apply, but those laws were not written with AI in mind and may require change and supplementation to serve as effective vehicles for accountability.

---

decisions, or recommendations. But one thing is for sure: it's a marketing term. Right now it's a hot one." *See* Michael Atleson, *Keep Your AI Claims in Check*, Federal Trade Commission (Feb. 27, 2023), https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

- Government procurement laws and policies for the acquisition of AI systems and services can provide a model and drive development of best practices.

As an illustrative example, we then analyze how these accountability mechanisms work in practice and could be improved in the particular context of AI in hiring and employment, in which CDT has extensive internal expertise and has published standards for civil rights accountability.

We close by discussing accountability for harms that result from content produced by generative AI. Generative AI tools may be used by individuals for a wide range of creative and expressive purposes, at least some of which are protected by the First Amendment. Generative AI tools are developed and used over multiple phases and by multiple types of actors, which makes accountability complicated. Nevertheless, each of the four accountability tools described above apply to generative AI. There likely is not a one-size-fits-all liability model that adequately protects individuals' rights in relation to generative AI tools, but it will be vital to map how existing legal principles across criminal and civil law would apply to cases involving generative AI to ensure that we do not end up with a liability gap that leaves serious threats to individuals' rights unaddressed. Policymakers should pursue accountability frameworks that focus on risk assessment and mitigation and incentivize the implementation of safeguards against abuse.

I. AI harms and the need for accountability

One of the key goals of accountability mechanisms should be to minimize the risk of harm to people and to provide for responsibility for such harms when they do occur. In this section, we briefly survey the types of harms that can result from the use of AI in three contexts: automated decision-making; systems for content analysis, moderation and recommendation; and generative AI. A key metric on which to evaluate whether accountability measures are adequate and meaningful is the extent to which they address these types of harms.

A. Automated decision-making

As CDT's March 2023 comments to the NTIA describe, automated decision-making plays a tremendous role in both the public and private sector.[13] In this section, we highlight several examples where either public agencies or private actors have deployed AI tools to perform functions that had traditionally been performed by humans. Although such process changes can have benefits such as greater efficiencies, they also can lead to harms such as discriminatory denials of access to economic opportunity.

1. Public sector uses & harms

Public agencies are increasingly using data and technology, including artificial intelligence, in their efforts to modernize service delivery. Unfortunately, the enthusiasm to use emerging technologies and products often outpaces public administrators' understanding and abilities to identify and mitigate bias, inequities, and other harms. The stories are numerous, and CDT briefly discusses three examples of public sector actors that have deployed artificial intelligence to deliver services and, in the process, have inflicted harm on the populations they aim to help: in education, in disability services, and in public housing.

Education: Endangering Students to Keep Them Safe

K-12 educational agencies and institutions are navigating a growing market of algorithmic decision-making systems designed to transform district and school functions such as assigning students to schools, preventing dropout, and keeping students safe. These decisions can significantly affect students' experiences, relationships, and future opportunities, whether by determining which school a student attends and thus what teachers and extracurriculars are available to her, or by deciding whether or not that student is a threat to school safety.[14]

[13] Ridhi Shetty, *et al*., *CDT Comments to NTIA on Privacy, Equity, and Civil Rights,* Center for Democracy & Technology (Mar. 6, 2023), https://cdt.org/insights/cdt-comments-to-ntia-on-privacy-equity-and-civil-rights/.
[14] Hannah Quay-de la Vallee & Natasha Duarte, *Algorithmic Systems in Education: Incorporating Equity and Fairness When Using Student Data*, Center for Democracy & Technology (Aug. 8, 2019), https://cdt.org/wp-content/uploads/2019/08/2019-08-08-Digital-Decision-making-Brief-FINAL.pdf.

Take student activity monitoring software, for example. As outlined in a recent letter to the U.S. Department of Education,[15] student activity monitoring software utilizes artificial intelligence to scan student messages and documents stored online or on school devices.[16] The resulting surveillance is pervasive: 89 percent of teachers report that their school uses student activity monitoring software, and monitoring often occurs outside of school hours.[17] Research reveals how online monitoring practices disproportionately harm protected classes of students and may violate civil rights laws:[18]

- **Exacerbating disproportionate discipline and law enforcement interactions for students of color.** As a result of student activity monitoring, students of color are experiencing increased interactions with law enforcement, as well as being disciplined at disproportionate rates. 44 percent of teachers report that students were contacted by law enforcement because of behaviors flagged by student activity monitoring. Moreover, 78 percent of teachers report that student activity monitoring flagged students for violations of disciplinary policy, and 59 percent report that a student was actually disciplined following those alerts. That discipline falls disproportionately along racial lines, with 48 percent of Black students and 55 percent of Hispanic students reporting that they or someone they know got into trouble as a result of student activity monitoring — compared to 41 percent of white students.

- **Targeting LGBTQI+ students for "outing," discipline, and criminal investigations.** LGBTQI+ students are disproportionately targeted as a result of student activity monitoring. 29 percent of LGBTQI+ students report that they or another student they know has had their sexual orientation or gender identity disclosed without their consent

[15] Letter Re: Discriminatory Effects of Online Monitoring of Students on LGBTQI+ Students, Students of Color, and Students with Disabilities, (Aug. 2, 2022), https://cdt.org/wp-content/uploads/2022/08/OCR-Letter-Final-August-2022.pdf
[16] *See* Hugh Grant-Chapman, *et al*., *Student Activity Monitoring Software: Research Insights and Recommendations*, Center for Democracy & Technology at 2 (Sept. 21, 2021), https://cdt.org/insights/student-activity-monitoring-software-research-insights-and-recommendations/.
[17] Elizabeth Laird, *et al*., *Hidden Harms: The Misleading Promise of Monitoring Students Online* at 8, Center for Democracy & Technology (Aug. 2022), https://cdt.org/wp-content/uploads/2022/08/Hidden-Harms-The-Misleading-Promise-of-Monitoring-Students-Online-Research-Report-Final-Accessible.pdf.
[18] *See Id.* at 19-24.

(i.e., "outed') due to student activity monitoring. Additionally, 56 percent of LGBTQI+ students reported that they or someone they know was disciplined as a result of student activity monitoring, and 31 percent reported they were contacted by law enforcement regarding a crime flagged by the software — compared to 44 percent and 19 percent, respectively, for their non-LGBTQI+ peers.

- **Harming students' expression and mental health.** Research also suggests that students with disabilities are experiencing disproportionate harm. Approximately five in ten students agree with the statement: "I do not share my true thoughts or ideas because I know what I do online may be monitored." This chilling effect is compounded for students with learning differences and physical disabilities, with 60 percent and 67 percent, respectively, reporting that they do not share their true thoughts or feelings due to monitoring. Moreover, 66 percent of teachers are concerned that students are less likely to access resources or visit websites that might provide help to them, such as how to share their sexual orientation or gender identities with their families or how to access mental health support.

Student activity monitoring, powered by AI, is subjecting protected classes of students to increased discipline and interactions with law enforcement, invading their privacy, and creating hostile environments for students to express their true thoughts and authentic identities. At minimum, this environment causes disparate impact[19] and—to the extent that monitoring software is expressly coded to flag words related to protected classes—may constitute disparate treatment.[20]

[19] *See* Alejandra Caraballo, *Remote Learning Accidentally Introduced a New Danger for LGBTQ Students*, Slate (Feb. 24, 2022), https://slate.com/technology/2022/02/remote-learning-danger-lgbtq-students.html#:~:text=Last%20year%2C%20a%20student%20in,becoming%20more%20and%20more%20invasive; *see also* Avery Kleinman, *Remote Learning Ushered in a New Era of Online Academic Surveillance. What's Next?*, 1A (Jan. 11, 2022), https://the1a.org/segments/remote-learning-ushered-in-a-new-era-of-online-academic-surveillance-whats-next/.
[20] *See Guardians Ass'n v. Civil Serv. Comm'n*, 463 U.S. 582, 607–08 (1983); *see also Cannon v. University of Chicago*, 441 U. S. 677, 691 (1979).

CENTER FOR DEMOCRACY & TECHNOLOGY

Disability Services: Improving Efficiency By Sacrificing Benefits Access

Government agencies are increasingly turning to algorithms to determine whether and to what extent people should receive crucial benefits through programs like Medicaid, Medicare, unemployment, and Social Security Disability. Billed as a way to increase efficiency and root out fraud, these algorithm-driven decision-making tools are often implemented without much public debate and are incredibly difficult to understand once deployed for use. Reports from people on the ground confirm that the tools are frequently reducing and denying benefits, often with unfair and inhumane results. Moreover, people with disabilities experience disproportionate and particular harm because of unjust algorithm-driven decision-making.[21]

An increasing number of states are turning to more automated algorithm-driven assessment and decision-making, relying on tools that quickly process multiple data inputs to evaluate whether a person needs assistance and what benefits they should receive. These tools also may be used to flag benefits recipients who appear to be defrauding the system, or used in the context of health care,[22] to determine how to distribute funding based on the type and amount of care some people should receive.[23]

As one example, state governments have adopted algorithm-driven decision-making to assess disabled people's eligibility for home- and community-based services (HCBS) under Medicaid. An Idaho budget allocation tool subject to ongoing litigation serves as a useful example of how

---

[21] Lydia Brown, *et al.*, *Challenging the Use of Algorithm-Driven Decision-Making in Benefits Determinations Affecting People with Disabilities*, Center for Democracy & Technology (Oct. 2020), https://cdt.org/wp-content/uploads/2020/10/2020-10-21-Challenging-the-Use-of-Algorithm-driven-Decision-making-in-Benefits-Determinations-Affecting-People-with-Disabilities.pdf.

[22] *See* Michele Gilman, *Did a Failed Algorithm Drive Welfare Recipients To Suicide?*, The National Interest (Feb. 18, 2020), https://nationalinterest.org/blog/buzz/did-failed-algorithm-drive-welfare-recipients-suicide-124691; *see also*, Michele Gilman, *AI Algorithms Intended to Root Out Welfare Fraud Often End Up Punishing the Poor Instead*, The Conversation (Feb. 14, 2020), https://theconversation.com/ai-algorithms-intended-to-root-out-welfare-fraud-often-end-up-punishing-the-poor-instead-131625 (reporting on Michigan's MiDAS system, which in 2013 made roughly 48,000 fraud accusations against unemployment insurance recipients – a five-fold increase from the prior system. A state review later determined that 93% of the fraud determinations were wrong).

[23] Eliza Strickland, *Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients*, IEEE Spectrum (Oct. 24, 2019), https://spectrum.ieee.org/racial-bias-found-in-algorithms-that-determine-health-care-for-millions-of-patients.

states may deploy such tools.[24] In 2011, Idaho adopted a new program to assess recipients' approved budgets for HCBS under Medicaid. Under the program, a person would travel to a medical assessment center where an Independent Assessment Provider (IAP) would complete a proprietary form that scored the person's need for assistance in feeding, toileting, dressing, and other functions. The IAP would manually enter that data into a digital Budget Tool, which, in turn, automatically calculated an Assigned Budget Amount for those reported needs based on data held in a proprietary database. An Assigned Budget Amount could only be increased if program managers found that the person required it for their "health and safety" – an undefined term that led to significant cuts to people's individualized budgets and lengthy and difficult appeals.

A closer look at Idaho's tool reflected substantially flawed design and execution. At trial, the judge found that the Budget Tool was developed based on a small, unrepresentative data set.[25] Additionally, IAPs had to record and then transfer large quantities of data, resulting in what the judge called a "high likelihood of human error." Although Idaho knew that the Budget Tool needed to be recalibrated annually to appropriately assess current costs, Idaho did not do that. The state agency did not provide people with a copy of the proprietary assessment form or allow them to access all of the form or its results. And Idaho had no process in place to audit whether budgets assigned by the tool accurately met peoples' needs.

Problems such as these illustrate the potential harmful effects of algorithm-driven decision-making tools, and the need for careful oversight and accountability for errors and resulting harms. Systems that rely on algorithm-driven needs assessments often make it challenging for beneficiaries to adequately challenge those decisions. The results can be devastating for their independence and quality of life.[26]

Public Housing: Using Technology to Surveil and Evict Tenants

Like many sectors, public housing agencies are also turning to data and technology in order to work more quickly and effectively. Funded and overseen nationally by the Department of

[24] *See K.W. v. Armstrong*, 180 F.Supp.3d 703 (D. Idaho, 2016).
[25] *Id*. at 714-16
[26] *See generally*, Lydia Brown *et. al*., *supra* at n. 21.

Housing and Urban Development (HUD), public housing agencies have a variety of responsibilities with the overarching goal of ensuring access to safe and affordable housing.

Public housing agencies are responsible for a large variety of regulations, including:

- Enforcing rules and regulations related to anti-discrimination (on the basis of race, disability status, and, in many localities, income), landlord harassment, and building code violations.
- Providing benefits to those who are experiencing housing instability, such as providing temporary housing or financial assistance.

However, using data and technology also present risks. One potential risk is that data-driven techniques, including AI, introduce unintended bias. Another risk is that the provision of technology services by third parties could result in the misuse of data and a loss of trust in public housing agencies.[27]

For example, recent reporting revealed that local public housing agencies are using federal grants administered by HUD to install surveillance camera systems that use facial recognition technology.[28] They claim that this technology is important to keep residents safe; however, agencies are using this technology to surveil residents, identify any who purportedly violated lease terms, and use this information to punish or evict residents from public housing. Evictions can have long-term consequences, including not being able to find alternative housing or even preventing individuals from gaining employment. Additionally, public housing is more heavily surveilled than other areas, and its residents are disproportionately Black and brown people. As a result, the adoption of this technology is most likely to punish people of color. This is another example of why public agencies, and the private companies that offer services on their behalf, need greater accountability when it comes to their uses of AI.

---

[27] Michael Yang, *The Promise and Peril of Data & Technology Use By Public Housing Agencies*, Center for Democracy & Technology (Mar. 2023), https://cdt.org/insights/report-the-promise-and-peril-of-data-technology-use-by-public-housing-agencies/.

[28] Douglas MacMillan, *Eyes On the Poor: Cameras, Facial Recognition Watch Over Public Housing*, Washington Post (May 16, 2023), https://www.washingtonpost.com/business/2023/05/16/surveillance-cameras-public-housing/.

2. Private sector uses & harms

In the private sector, data- and algorithm-driven processes are influencing—if not determining outright—who can get and maintain access to key economic opportunities such as housing, financial support, and employment. When AI systems are deployed in these high-risk settings without responsible design and accountability measures, it can devastate people's lives.

Targeted behavioral advertising can affect who even learns of the availability of a housing, credit, or job opportunity or certain services. This process involves using people's online data and activity to identify groups whose characteristics appear to correlate more often with interest or eligibility for an advertised opportunity, and delivering advertisements for the available opportunity to those specified audiences.[29] Target audiences can end up excluding people whose communities have historically encountered barriers to these opportunities, because their characteristics differ from those who have been considered eligible or preferred candidates for housing, employment, and credit.[30] As a result, targeted advertising can contribute to disparities in accessing critical opportunities by reinforcing a cycle of directing more advertisements for these opportunities to the groups that already have greater access to them.

When a person does learn of an available housing, credit, or job opportunity and proceeds to apply, automated decision-making can impact the person's chances of being selected for that opportunity. Renters are affected by automated processes used in both the public and private sector particularly for two purposes: to screen rental applicants and to monitor tenants. Landlords will screen rental applicants to predict risks of nonpayment of rent or legal liability for threats to other tenants or property.[31] Automated screening processes can disqualify applicants with little explanation based on credit, criminal, and eviction records that are often unreliable

[29] Jinyan Zang, *Solving the Problem of Racially Discriminatory Advertising on Facebook*, Brookings Institution (Oct. 19, 2021), https://www.brookings.edu/research/solving-the-problem-of-racially-discriminatory-advertising-on-facebook/.
[30] Tanya Kant, *Identity, Advertising, and Algorithmic Targeting: Or How (Not) to Target Your "Ideal User,"* MIT Schwarzman College of Computing (2021), https://mit-serc.pubpub.org/pub/identity-advertising-and-algorithmic-targeting/release/2.
[31] Tex Pasley, Henry Oostrom-Shah, & Eric Sirota, *Screened Out: How Tenant Screening Reports Undermine Fair Housing Laws and Deprive Tenants of Equal Access to Housing*, Urban Institute (2021), https://www.povertylaw.org/wp-content/uploads/2021/01/tenant-screening-final-report.pdf.

for making such predictions and they are most likely to disqualify rental applicants from communities that experience higher rates of socioeconomic marginalization such as Black and immigrant communities, disabled people, and domestic violence survivors.[32] Surveillance technologies are more likely to misidentify and incorrectly flag people of color and people with disabilities as threats, and the use of these technologies in housing can trigger law enforcement interactions that contribute to criminal record data that will be considered in future tenant screening.[33] Similar to rental applicants, the credit and court record data used to evaluate loan applicants can limit access to mortgage loans or subject certain student borrowers to higher interest rates.[34]

Workers can encounter automated decision-making tools at any point of the job cycle, from application to interview to promotion. Well-known examples of these tools perform resume screening, administer scored questionnaires or gamified assessments, and analyze recorded video interviews.[35] In many cases, these tools are created by analyzing "successful" employees to identify traits for which future candidates are then assessed. The risks in this approach are obvious: if the data used to train the AI system is not representative of wider society or reflects historical patterns of discrimination, it can reinforce existing bias and lack of representation in the workplace.[36]

---

[32] *See* Center for Democracy & Technology, Comments to Federal Agencies on Tenant Screening (June 2, 2023), https://cdt.org/insights/cdt-comments-to-federal-agencies-highlight-risks-of-data-used-in-tenant-screening/; *see also*, Lydia X. Z. Brown, *Tenant Screening Algorithms Enable Racial and Disability Discrimination at Scale, and Contribute to Broader Patterns of Injustice*, Center for Democracy & Technology (July 7, 2021), https://cdt.org/insights/tenant-screening-algorithms-enable-racial-and-disability-discrimination-at-scale-and-contribute-to-broader-patterns-of-injustice/.

[33] Nick Klepper, *Meet the Spy Tech Companies Helping Landlords Evict People*, Vice: Motherboard (Jan. 4, 2023), https://www.vice.com/en/article/xgy9k3/meet-the-spy-tech-companies-helping-landlords-evict-people.

[34] Center for Democracy & Technology, Comments to Financial Regulators on Financial Institutions' Use of Artificial Intelligence (Jul. 1, 2021), https://cdt.org/wp-content/uploads/2021/07/2021-07-01-CDT-Request-for-Information-and-Comment-on-Financial-Institutions-Use-of-Artificial-Intelligence-including-Machine-Learning.pdf.

[35] Lydia X.Z. Brown, Ridhi Shetty, & Michelle Richardson, *Algorithm-Driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?,* Center for Democracy & Technology (Dec. 3, 2020), https://cdt.org/insights/report-algorithm-driven-hiring-tools-innovative-recruitment-or-expedited-disability-discrimination/.

[36] *See, e.g.*, Keith E. Sonderling, Bradford J. Kelley, & Lance Casimir, *The Promise and the Peril: Artificial Intelligence and Employment Discrimination*, 77 U. Miami L. Rev. 1 (2022), https://repository.law.miami.edu/umlr/vol77/iss1/3.

Moreover, these tools often claim to predict personality traits, aptitudes, and skills that employers may desire, but that may not be necessary to perform essential job functions. For example, a resume might indicate a person's participation in certain affinity groups and lack extracurriculars that are not affordable or widely accessible – even if the former experience contributed to the person's professional development, this resume might be disqualified if the latter experience is among factors that an employer tends to favor. Even when any of these qualities are necessary for particular roles, the data being measured is not always the most reliable or only reliable predictor of job performance. People with disabilities that affect how they demonstrate desired skills or traits can be rejected because they do not conform to the specific ways in which previously selected candidates demonstrated the same qualities.

Automated decision-making also extends to the workplace, as employers use electronic surveillance and algorithmic management tools to monitor workers' performance, productivity, movements, and interactions with coworkers and customers to make decisions about disciplinary actions, compensation, promotion, or termination.[37] Such surveillance has been shown to harm workers' mental and physical wellbeing, hinder their ability to organize, and limit their access to rights of employment.

B. Uses and harms in the online content space

Public and private sector actors alike rely on automated content analysis for a range of purposes. This category includes natural language processing (NLP) tools,[38] including large language models,[39] as well as techniques from the field of computer vision that enable the automated evaluation of images and video.[40] Different machine learning techniques for analyzing the meaning of digital content can be useful in a variety of applications, including

[37] Matt Scherer, *CDT, GFI, Others Send Memos Urging White House to Take Action on Electronic Workplace Surveillance*, Center for Democracy & Technology (Apr. 3, 2023),https://cdt.org/insights/cdt-gfi-others-send-memos-urging-white-house-to-take-action-on-electronic-workplace-surveillance/.
[38] Natasha Duarte, Emma Llansó, & Anna Loup, *Mixed Messages: The Limits of Automated Social Media Content Analysis*, Center for Democracy & Technology (Nov. 28, 2017), https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/.
[39] Nicholas & Bhatia, *supra* n. 4.
[40] Carey Shenkman, Dhanaraj Thakur, & Emma Llansó, *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*, Center for Democracy & Technology (May 20, 2021), https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/.

content moderation, automated decision-making, and recommendation systems. But these tools also have real limitations that can create significant risks to people's rights.

For example, automated content analysis tools are commonly used by the online service providers that host, link to, or otherwise facilitate access to user-generated content as part of their content moderation systems. Many of these tools rely on machine learning classifiers that provide a probabilistic assessment of the likelihood that a given post or file uploaded by a user meets a certain condition, such as "contains nudity" or "is hate speech." Providers may also use perceptual hashing tools to identify and automatically block images or videos that the provider has previously seen and removed from its service.[41] While this kind of automated enforcement may pose a lower risk to fundamental rights in cases where the target content (such as child sexual abuse material (CSAM)) is always prohibited regardless of its context, for many types of content, meaning is highly dependent on context: the same video of a terror attack could represent praise for the terrorists and incitement to others to commit similar acts, news reporting, or efforts by human rights advocates to document atrocities for future legal action.[42] Automated removal of content in such circumstances can significantly restrict speech on challenging and important topics.

Moreover, the efficacy of machine learning tools is highly influenced by the data they have trained on. For tools parsing human communication, this can lead to significant disparities in, for example, whose speech is deemed "toxic," when a tool is used to analyze speech that is from a different community of speakers than those represented in its training data.[43] It is also important to recognize that the development of language analysis capabilities in languages other than English lags far behind English-language NLP, due to disparities in the available digitized text resources to train NLP models as well as in the investments in research and

[41] Perceptual hashing tools allow for the creation of a "digital fingerprint" of a file that is somewhat robust to different manipulations or alterations of that file (in contrast to cryptographic hashes that are designed to change if the underlying file varies by a single bit). This hash can then be compared to the hash of a newly uploaded file to identify a likely match. *Id.* at 13.

[42] Belkis Wille, *Video Unavailable: Social Media Platforms Remove Evidence of War Crimes*, Human Rights Watch (Sep. 10, 2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.

[43] *See* Alessandra Gomes, Dennys Antonialli, & Thiago Oliva, *Drag Queens and Artificial Intelligence: Should Computers Decide What Is 'Toxic' on the Internet?*, INTERNET LAB (June 28, 2019), https://internetlab.org.br/en/news/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/.

development in both the public and private sectors.[44] Some researchers and online services are experimenting with the development of multilingual large language models, which promise to be able to analyze a less-common language based on only a fraction of the sample text used to train English-language models. But there are limits to multilingual LLMs' ability to accurately evaluate meaning in diverse languages, and without research and additional transparency into the training and efficacy of these models, there is a risk that language analysis in non-English languages flattens linguistic diversity, misses important cultural context, and ultimately gives second-class treatment to languages other than English.[45] In practice, this could result in disproportionate removal or restriction of non-English content, including both over-removal of innocuous speech and under-removal of truly threatening, harassing, or otherwise abusive content that the models cannot recognize.

Automated content analysis tools are not only used in the online content moderation context. They may be incorporated into automated decision making systems for everything from filtering school children's internet access[46] to evaluating resumes[47] to assessing whether an individual should be granted a visa to enter the United States.[48] In short, automated content analysis tools are likely at use any time a public or private entity is attempting to evaluate a large body of content, meaning that the limitations of these tools have the potential to reverberate throughout many aspects of people's lives, particularly if these tools are used to automatically enforce decisions against individuals. For example, limitations on non-English NLP models mean that people could be denied public benefits or political asylum because the automated tool used to recommend, or even make, a decision could not accurately analyze an application written in a non-English language. Thus, it is vital that both the creators of these tools, and the public- and private-sector entities that make use of them, provide significant transparency into the design, operation, and use of any machine learning tools that incorporate automated content analysis features.

---

[44] *See* Nicholas & Bhatia, *supra* n. 4.
[45] *Id.*
[46] *See* Grant-Chapman, *et al.*, *supra* n. 16.
[47] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
[48] *Coalition Letter Urges the Department of Homeland Security to End Social Media Monitoring Initiative*, Brennan Center for Justice (Aug. 24, 2021), https://www.brennancenter.org/our-work/research-reports/coalition-letter-urges-department-homeland-security-end-social-media.

Beyond content analysis and decisions to remove content from an online service, providers also employ automated decision-making in the form of content recommendation algorithms. Ranking and recommendation algorithms are central to the organizing and distribution of content and can provide significant utility to individuals who are seeking information online.[49] When they operate opaquely, however, they can lead people across the political spectrum and from all walks of life to suspect that they have been "shadowbanned," i.e. had the distribution of their content intentionally restricted in a non-transparent way.[50] When users are kept in the dark about this practice, it can lead to conspiracy theories and an overall lack of trust in online services. It also prevents users from understanding and conforming to the rules of a particular service so that they may access a broader audience.[51]

C. Generative AI

While the so-called "existential risks" of generative AI models, and highly-capable models more broadly, have garnered significant media attention following public letters from tech company leaders,[52] the immediate, documented risks and harms arising from generative AI demand immediate attention. We therefore focus on these risks and harms, while recognizing the importance of thinking through longer-term implications of these new technologies. Moreover, there is reason to believe that addressing the immediate concerns detailed below may help mitigate "existential risks," or at least sharpen the field's collective analysis of the necessary tools and interventions.

---

[49] Caitlin Vogus, Emma Llansó, & Samir Jain, *CDT and Technologists File SCOTUS Brief Urging Court to Hold that Section 230 Applies to Recommendations of Content*, Center for Democracy & Technology (Jan. 18, 2023), https://cdt.org/insights/cdt-and-technologists-file-scotus-brief-urging-court-to-hold-that-section-230-applies-to-recommendations-of-content/.

[50] *See* Gabriel Nicholas, *Shedding Light on Shadowbanning*, Center for Democracy & Technology (Apr. 26, 2022), https://cdt.org/insights/shedding-light-on-shadowbanning/.

[51] *Id.*

[52] *See Pause Giant AI Experiments: An Open Letter*, Future of Life Institute (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/; *see also Statement on AI Risk*, Center for AI Safety https://www.safe.ai/statement-on-ai-risk#open-letter.

1. Equity

As with most AI systems, equity is a concern. Generative AI systems are trained on data that reflect the biases of the world that data stems from. This, in turn, can lead to those biases and prejudices becoming embedded in the AI itself.[53] While designers can take steps to limit this bias on both the input end (by curating the training data and trying to eliminate biases at that stage) and on the output end (trying to detect outputs that reflect a bias and stopping or modifying those outputs before they go to the user), neither of these approaches will be completely effective.[54] Because of the enormous volume of data needed to train generative AI systems, it is typically infeasible to have humans vet all the training data. Additionally, if the system is designed to continue learning from user queries and responses over time, those inputs will generally be outside the control of the system developers.

Bias embedded in generative AI systems is particularly concerning in an education context where students may be using these tools to learn more about the world around them, meaning the tool may impart or reinforce biases in students' thinking.[55] And as discussed earlier, most large language models are trained on English-language data, and perform less well in other languages, raising serious and pervasive equity concerns.[56]

2. Privacy

Generative AI systems also raise privacy risks. On one hand, there is the question of how training data is sourced. Personal information about individuals and content they have created is being used to train AI systems. If the generative AI system uses existing data corpuses that are restricted to the purpose of training AI systems, this may be less of a concern. However, any

---

[53] *See* Kieran Snyder, *ChatGPT Writes Valentines*, Textio (Feb. 14, 2023), https://textio.com/blog/chatgpt-writes-valentines/102332725392. *See also* Leonardo Nicoletti and Dina Bass, *Humans Are Biased. Generative AI Is Even Worse*., Bloomberg Technology + Equality (June 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[54] Davey Alba, OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails, Bloomberg Equality (Dec. 8, 2022), https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results.

[55] Vinitha Gadiraju, *et al.*, *"I Wouldn't Say Offensive But…": Disability-Centered Perspectives on Large Language Models*, Proceedings of FAccT (2023), https://research.google/pubs/pub52358/.

[56] *See* Nicholas & Bhatia, *supra* n. 4.

system that gathers data from the public internet for novel purposes may be using data in ways that the data subjects did not anticipate and may not be comfortable with, even if they technically consented to it under a broad clause enabling unspecified future uses. Moreover, there is also a risk of this personal information being revealed as a result of user prompts or through prompt attacks.[57]

A further privacy risk arises with respect to the data that is inputted or created during interaction with the system, whether that be the outputs from the data or prompts and queries provided to the system. In some ways, this mirrors existing concerns around search engine privacy, though it may be exacerbated to the extent that the design of chatbots encourages more detailed or intimate "conversations" than mere entry of a search term. For example, a student asking for resources around gender and sexuality may be placed at risk if their teachers or school administrators get access to these queries and the student is outed to their family and community.

3. Efficacy and Accuracy

Another critical concern with generative AI systems is one of efficacy—in other words, whether they actually work as intended or advertised. Because of the unsupervised nature of their development, generative systems may "hallucinate," meaning they generate untrue responses.[58] Whether this is a problem depends on how the user is interacting with the system. If they asked the system to write a short fictional story, then untruth is not an issue; in fact it is expected. However, if the user was asking a factual question for research or for purposes of seeking medical information, a hallucination is a failure case that could even lead to physical harm if a user acts on the information provided. It can be difficult for system developers to address this issue. This is both because they may not wish to restrict the system from hallucinating entirely, because, for example, it is a multi-use system where creativity may be desirable, and because it simply may not be possible. These systems are language models—that is, they are creating a model of how words, phrases, and sentences are typically associated with each other across an

[57] Nicholas Carlini, *et al*., *Extracting Training Data from Large Language Models* (Jun. 15, 2021), https://arxiv.org/pdf/2012.07805.pdf.
[58] Craig G. Smith, *Hallucinations Could Blunt ChatGPT's Success*, IEEE Spectrum (Mar. 13, 2023), https://spectrum.ieee.org/ai-hallucination.

enormous corpus of text, so that they can predict and generate text that will be deemed responsive to a given prompt. This process enables these models to produce grammatically correct text that often aligns true, factual information, but the model does not "learn" ground truth or falsity, and an untrue statement or hallucination may be an equally adequate prediction as a true statement.[59]

4. Fraud and Deepfakes

Generative AI tools also are likely to exacerbate fraud, as tools make it easier to quickly generate massive amounts of convincing text for span or phishing, as well as personalized scams, or to trick people by impersonating a familiar voice.[60] Deepfakes – videos or images that have been digitally manipulated to misrepresent the voice and likeness of another person – can misrepresent public figures or events in a way that threatens elections, national security, and general public order.[61] Just last month, AI-generated images depicting explosions at the Pentagon went viral on social media, which some news outlets reported as "breaking news" even though the explosions never happened.[62] Deepfakes can also be used to defraud, harass,

[59] This may be especially true if the data a model is trained on contains a high volume of false statements ("the Earth is flat") and encodes those associations (between "Earth" and "flat") strongly.

[60] Steve Mollman, *Scammers are Using Voice-Cloning A.I. Tools to Sound Like Victims' Relatives in Desperate Need of Financial Help. It's Working*, Fortune (Mar. 5, 2023), https://fortune.com/2023/03/05/scammers-ai-voice-cloning-tricking-victims-sound-like-relatives-needing-money/.

[61] *See* Shannon Bond, *Fake Viral Images of an Explosion at the Pentagon Were Probably Created by AI*, NPR (May 22, 2023), https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai; *see also*, David Klepper & Ali Swenson, *AI Presents Political Peril for 2024 with Threat to Mislead Voters*, AP News (May 14, 2023), https://apnews.com/article/artificial-intelligence-misinformation-deepfakes-2024-election-trump-59fb51002661ac5290089060b3ae39a0 .

[62] *See Is It Real or Is It AI? Indian News Outlets Run Fake Image of Pentagon Fire as 'Breaking News,'* News Laundry (May 23, 2023), https://www.newslaundry.com/2023/05/23/is-it-real-or-is-it-ai-indian-news-outlets-run-fake-image-of-pentagon-fire-as-breaking-news; *see also* Chloe Xiang, *Verified Twitter Accounts Spread AI-Generated Hoax of Pentagon Explosion*, Vice (May 22, 2023), https://www.vice.com/en/article/7kx84b/ai-generated-pentagon-explosion-hoax-twitter.

and extort people.[63] None of these harms is new, but they are made cheaper, faster, and more effective by the ease, speed and widespread accessibility of generative AI tools.

For example, in previous elections, operatives used robocalls to spread incorrect information about mail-in voting in an effort to suppress Black voter turnout,[64] and deceptive text messages to spread intentionally misleading voting instructions for a Kansas ballot initiative in 2022.[65] Bad actors could use AI to exponentially grow and personalize voter suppression or other targeting efforts, increasing their harmful impact. Today, consumers can often spot a scam email, text or robocall because it uses non-personalized language and there may be grammatical or language errors (or, in the case of robocalls, a notably automated voice). Generative AI tools will make it easier to create tailored, accurate, realistic messages that draw victims in.

More generally, the growth of inauthentic content makes it harder for people to know what news and content they can trust, such that even authentic content is undermined. Journalists, whistleblowers, and human rights defenders are experiencing these effects already, facing higher hurdles than ever before to establish and defend their credibility.[66]

II. Accountability for AI Systems

Given the range of harms that AI can cause, accountability is a critical piece of AI governance. It can help minimize harms, allocate responsibility and liability when harms do occur, and help to engender trust in the use and deployment of AI.

---

[63] *See e.g.*, Henry Ajder, Giorgio Patrini and Francesco Cavalli, *Automating Image Abuse: Deepfake Bots on Telegram*, Sensity, (Oct. 2020) (deepfake bots on Telegram digitally "undress" more than 100,000 women on the platform) https://www.medianama.com/wp-content/uploads/Sensity-AutomatingImageAbuse.pdf; *see also*, Thomas Brewster, *Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find*, Forbes (Oct. 14, 2021), https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/ (audio deepfake of executives' voices used to steal millions of dollars from companies).

[64] Christine Chung, *They Used Robocalls to Suppress Black Votes. Now They Have to Register Voters*, The New York Times (Dec. 1, 2022), https://www.nytimes.com/2022/12/01/us/politics/wohl-burkman-voter-suppression-ohio.html.

[65] Isaac Stanley-Becker, *Misleading Kansas Abortion Texts Linked to Republican-Aligned Firm*, The Washington Post, (Aug. 2, 2022), https://www.washingtonpost.com/politics/2022/08/02/kansas-abortion-texts/.

[66] Sam Gregory, *Tracing Trust: Why We Must Build Authenticity Infrastructure that Works for All*, Witness (May 2020), https://blog.witness.org/2020/05/authenticity-infrastructure/.

A. The accountability toolbox

No single tool will be sufficient to create accountability for AI systems. Here, we discuss four key tools: transparency, explainability, and interpretability; audits and assessments; laws and liability; and government procurement rules. These tools are mutually reinforcing. For example, a publicly released audit provides a measure of transparency, while transparency provides information necessary to determine whether liability should be imposed.

1. Transparency, explainability, and interpretability

As NIST has explained, "[a]ccountability presupposes transparency," and transparency in turn "reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system."[67] Explainability describes "'how' a decision was made in the system," while interpretability answers "'why' a decision was made by the system and its meaning or context to the user." The three are "distinct characteristics that support each other."[68]

The first hurdle on the quest for AI accountability is knowing not only that a harm occurred, but that it is attributable to an AI system. This is a higher barrier than one might expect: many AI systems operate in the background, impacting people's lives in ways that are invisible yet profound. For example, when employers use automated tools to sift through resumes or to analyze video recordings of candidate interviews to narrow down their applicant pool, job-seekers may only learn whether they've been selected to move ahead in the recruitment process. They may not be told on what basis the decision has been made, how similarly situated candidates were treated, or whether AI was involved in the decision. Similarly, benefits recipients may only be told that their monthly allocation is being reduced, and lack necessary information to ascribe their diminished income to an automated decision-making system. And indeed, internet users currently have no way of knowing whether text and images they encounter online were created by a person, a generative AI tool, or some combination of the

[67] *Artificial Intelligence Risk Management Framework*, National Institute of Standards and Technology § 3.4 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[68] *Id*. § 3.5.

two, though, as discussed below, watermarking and other techniques may be able to provide such transparency in at least some cases.

Of course, knowing that AI was used is only an initial step. Accountability also requires disclosure of information such as how a system was trained and on what data sets, its intended uses, how it works and is structured, and other information that permits the intended audiences (which can include affected individuals, policymakers, researchers, and others) to understand how and why the system makes particular decisions. We describe below in more detail a set of standards that CDT developed in conjunction with civil rights groups that describes in more detail the types of information necessary for effective transparency in the specific context of automated employment decision tools.

Standardized formats for transparency have begun emerging, including "Datasheets for Datasets"[69] to document the datasets used to train machine learning models; model cards[70] that set out key parameters and characteristics of individual models; and system cards that some companies have begun to release that describe not only an individual model, but the broader system in which one or more models may be embedded. Such standardization can be a useful step because it can make it easier, particularly for users, to understand the information provided.

One context in which transparency around the use of automation is comparatively well developed is in the online content space, where it is well established that transparency is a necessary—though by no means sufficient[71]—prerequisite for accountability and respect for civil and human rights. Civil society organizations have long called on social media platforms to disclose the content rules that they expected their users to follow, the processes used to enforce those rules, and data about the nature and volume of content restriction actions taken by the platform.

---

[69] Timnit Gebru, *et al*., *Datasheets for Datasets*, (Dec. 1, 2021), http://arxiv.org/abs/1803.09010.
[70] Margaret Mitchell, *et al*., *Model Cards for Model Reporting,* (Jan. 14, 2019), https://arxiv.org/abs/1810.03993.
[71] Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 New Media & Society 3 at 973-989 (Dec. 13, 2016), https://journals.sagepub.com/doi/10.1177/1461444816676645.

As the use of algorithmic systems for content moderation and recommendation became widespread, expectations for AI transparency were encoded in normative documents like the Santa Clara Principles on Transparency and Accountability in Content Moderation (SCPs).[72] The SCPs call for improved transparency to users about when automation has been employed in decisions that remove a user's speech from the service or restrict its distribution, and urge companies to design their algorithmic systems with "explainability" in mind, so that it is possible to develop human-understandable explanations of how and why an algorithmic system reached a certain decision.[73] CDT has also discussed the need for improved transparency in the operation of ranking and recommendation algorithms, as opaque content moderation practices (colloquially referred to as "shadowbanning") undermine user trust in the information ecosystem and create fertile ground for conspiracy theories about viewpoint-based suppression of speech.[74] CDT recommends that companies adopt a co-design research methodology to determine what types of transparency will be most valuable and meaningful to their users.[75]

There are also significant opportunities to improve transparency and accountability of the operation of automated content analysis and recommendation systems by opening up more data to independent researchers. There is much that policymakers, civil society, journalists, and the general public need to understand about the operation of automated systems in shaping our information environment, and additional independent research is crucial to developing this understanding and guiding future policymaking.

Recently, state and federal legislators[76] have considered proposals to *mandate* certain forms of transparency from online service providers, including in the Florida and Texas state social media laws that are being challenged before the U.S. Supreme Court.[77] While mandatory transparency may serve some compelling state interests in fairness and accountability of how individuals'

---

[72] The Santa Clara Principles on Transparency and Accountability in Content Moderation, https://santaclaraprinciples.org/.
[73] *Id.*
[74] *See* Gabriel Nicholas, *Shedding Light on Shadowbanning, supra* at n. 50.
[75] Michal Luria, *"This is Transparency to Me": User Insights into Recommendation Algorithm Reporting*, Center for Democracy & Technology (Oct. 2022), https://cdt.org/wp-content/uploads/2022/10/algorithmic-transparency-ux-final-100322.pdf.
[76] Platform Accountability and Consumer Transparency Act of 2021, S. 797, 117th Cong. (2021) https://www.congress.gov/bill/117th-congress/senate-bill/797/text.
[77] *See NetChoice, LLC v. Moody*; *see also, NetChoice, LLC v. Paxton.*

speech is governed online, there are also significant First Amendment concerns about transparency mandates being leveraged by state actors to achieve particular content moderation outcomes.[78]

2. Audits & assessments

Algorithmic assessments and audits are a key component for evaluating AI-based systems to ensure accountability. We should normalize the principle that companies designing and deploying AI tools must identify potential risks, disclose the steps they have taken to mitigate those risks, and evaluate the effectiveness of those steps. Moreover, disclosure of such audits and assessments can also increase transparency.[79]

Audits and assessments can occur at different stages of the system life cycle and have different purposes.[80] *Pre-deployment assessments* put systems through a series of tests designed to determine how the system will behave in different scenarios in order to identify risks and evaluate potential mitigation measures. Examples include the following:

- *Auditing the training data.* This approach is intended to avoid embedding biases from the data into the system itself. Typically, an audit of training data is looking for ways in which historical biases and discrimination may lead to inaccurate or harmful outcomes.[81] If the data is found to be biased, developers must then determine how to correct those biases, either in the data itself or in the system overall.[82]

---

[78] *See* Brief for Center for Democracy & Technology, *et al*., as Amicus Curiae, *Twitter v. Paxton* (9th Cir., Apr. 11, 2022), *available at* https://cdt.org/wp-content/uploads/2022/04/59-CDT-Amicus-Brief-EFF-RSI-Support-Rehearing-in-Twitter-v-Paxton.pdf; *see also,* Daphne Keller, *Platform Transparency and the First Amendment* (Mar. 3, 2023) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4377578; *also see* Eric Goldman, *The Constitutionality of Mandating Editorial Transparency*, 73 Hastings Law Journal 1203 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005647.

[79] *See* Caitlin Vogus & Emma Llansó, *Making Transparency Meaningful: A Framework for Policymakers*, Center for Democracy & Technology (Dec. 14, 2021), https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers/.

[80] Inioluwa Deborah Raji, *et al*. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (Jan. 2020), https://dl.acm.org/doi/abs/10.1145/3351095.3372873.

[81] *See* Timnit Gebru, *et al*., *supra* at n. 69.

[82] Pedro Saleiro, *et al*., *Aequitas: A Bias and Fairness Audit Toolkit* (2018). https://arxiv.org/abs/1811.05577.

- *Assessing the development environment.* This approach focuses on whether the development environment has the correct frameworks and attributes to increase the likelihood of producing a system that minimizes bias and maximizes effectiveness, security, and other desirable characteristics.[83] This may mean building a diverse development team (across demographic aspects as well as expertise, including members with policy, legal, sociological, and ethical expertise, in addition to technical talent), incorporating assessment points throughout the development process to identify and correct issues early and often, and having well-supported avenues for developers to raise issues and concerns throughout the development process.

- *Human rights impact assessments.* These assessments are intended to identify potential impacts of an AI system on human rights ranging from privacy and non-discrimination to freedom of expression and association.[84] Such assessments can help identify, understand, and address potential adverse effects of the rights of affected stakeholders.

Pre-deployment audits and assessments are not sufficient because they may not fully capture a model or system's behavior after it is deployed and used in particular contexts.[85] Accordingly, *post-deployment audits and assessments*, those analyzing the system in situ after it is in use, interacting with consumers, are also critical. Broadly speaking, these can typically fall into two categories:

- *White-box assessments* are those that have full access to the inner workings of the system, and therefore are done by the developers or by third-parties with the involvement of the developers.[86] These assessments are typically able to be more

[83] Jacob Metcalf, *et al.*, *Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts*, Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency (2021), https://dl.acm.org/doi/abs/10.1145/3442188.3445935.

[84] Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, Facebook Newsroom (Nov. 5, 2018), https://www.researchgate.net/profile/Myo-Myanmar/publication/328842344_An_Independent_Assessment_of_the_Human_Rights_Impact_of_Facebook_in_Myanmar/links/5be60b53299bf1124fc77d27/An-Independent-Assessment-of-the-Human-Rights-Impact-of-Facebook-in-Myanmar.pdf.

[85] Joaquin Quinonero-Candela, *et al.*, eds. *Dataset Shift in Machine Learning*, (2008), *available at* https://books.google.com/books?hl=en&lr=&id=qJ0jEAAAQBAJ&oi=fnd&pg=PR9&dq=Dataset+Shift+in+Machine+Learning&ots=3wBXS3h7Jg&sig=oI9s-vBZgjj5SbviUb9eTN3AWvI#v=onepage&q=Dataset%20Shift%20in%20Machine%20Learning&f=false.

[86] *See* Inioluwa Deborah Raji, *et al., supra* at n. 80.

fulsome due to their increased access, but may be limited to what the developers and deployers flag as concerns.

- *Black-box assessments* are done by third parties without any access to the inner workings of the systems, focusing only on the outputs of the systems that are visible to consumers.[87] These audits can place some power back into the hands of consumers that are impacted by the system, but are limited by both their lack of access to the system, and their lack of power to make changes if concerns or issues are uncovered by the assessment.[88]

Although audits and assessments are a key piece of providing accountability, there remain fundamental questions about how to conduct them effectively.[89] A threshold question is *who* should conduct these types of audits and assessments. They can potentially be conducted by a wide range of actors: developers of algorithmic systems may conduct self-assessments as part of either private- or public-facing evaluations; independent researchers, civil society organizations, and independent experts[90] have conducted various evaluations and investigations into the operation of systems as part of broader accountability projects; users of a system with members of the general public may independently audit a system by monitoring everyday usage for bugs;[91] and professional auditing firms may conduct formal audits or assessments. This latter

[87] Christian Sandvig, *et al*., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, Data and Discrimination: Converting Critical Concerns into Productive Inquiry (2014), at 4349-4357, https://www.kevinhamilton.org/share/papers/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

[88] Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, AAAI/ACM Conf. on AI Ethics and Society (2019), https://dl.acm.org/doi/abs/10.1145/3306618.3314244.

[89] Sasha Costanza-Chock, Inioluwa Deborah Raji, & Joy Buolamwini, *Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem*, 2022 ACM Conference on Fairness, Accountability, and Transparency (2022), https://dl.acm.org/doi/abs/10.1145/3531146.3533213.

[90] Laura Murphy, *Facebook's Civil Rights Audit – Final Report* (Jul. 8, 2020) https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

[91] *See e.g.*, a discussion of bug bounties on Twitter found at Kyra Yee & Irene Font Peradejordi, *Sharing Learnings from the First Algorithmic Bias Bounty Challenge,* (Sept. 7, 2021), https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge; *see also*, a discussion of user-driven audits available at Alicia DeVos, *et al*., *Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (2022), https://dl.acm.org/doi/abs/10.1145/3491102.3517441.

category, formal audits by trained auditors, is still nascent in the field of algorithmic accountability.

This issue is already emerging in the European Union, where the new Digital Services Act (DSA) sets out an obligation for providers of very large online platforms to receive an independent audit, on at least an annual basis, to verify the provider's compliance with its obligations under the DSA.[92] These include an obligation to conduct a risk assessment of the provider's algorithmic systems and to mitigate the risks those services may pose to human rights. The European Commission is in the midst of developing the implementing legislation that defines the obligations of providers and auditors in carrying out these audits, but one of the challenges is how to ensure both independence and the necessary resources and expertise necessary to meaningfully evaluate the operation of online services' systems and processes, which will vary across different providers. Often, this technical expertise lies with the service provider itself, or individuals who have recently been employed by an online service provider but themselves are not clearly independent. Even auditing firms may be subject to capture by providers since providers may be reluctant to retain auditors that conduct truly independent and rigorous audits as compared to those who engage in more superficial exercises.

Experts in civil society, academia, and the technical community may have the necessary independence, but they will often lack the resources and capacity to engage in rigorous audits, particularly as AI systems become more widespread. Meaningful evaluations of company practice may require some combination of auditing firms with access to highly sensitive business information and broad consultations with independent experts who can contribute to audits and assessments and also serve as a check on their rigor and validity.

A second crucial issue is the development of standards to be used by auditors. The role of auditors is to evaluate company practice against an independent set of standards that articulate what the boundaries of "compliance" look like in a given regulatory regime—not to develop the standards themselves. Such standards will often embody policy and value judgments: standards for an audit designed to evaluate whether a system is biased, for example, may have to set forth

[92] Digital Services Act, Regulation (EU) 2022/2065, Article 37 (19 Oct. 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065&from=EN.

how much variation in performance, if any, is permissible across race, gender, or other lines in order to still be considered unbiased.

These standards can be built on top of principles developed through consultation with a wide range of stakeholders such as those embodied in the Administration's Blueprint for an AI Bill of Rights and NIST's AI Risk Management Framework (AI RMF).[93] The AI RMF can also be built upon through the development of "use profiles," which would explain how the AI RMF can be applied to specific sectors. Such profiles—if crafted carefully with input from groups affected by AI harms and advocates seeking to address these harms[94]—can be especially valuable because the different legal standards across sectors may necessitate different approaches in the development of standards in auditing.

Ultimately, any standards-development process must be open and multistakeholder to incorporate the necessary expertise in how algorithmic systems work, at a technical level, what kinds of risks they may pose to human rights and civil liberties, and input from affected stakeholders about mitigation of those risks. Auditing can be a useful tool for building trust in a technology ecosystem and in ensuring accountability, but only if it is approached in a rigorous, multistakeholder fashion.

3. Laws and liability

One of the key ways of ensuring accountability is the promulgation of laws and regulations that set standards for AI systems and impose potential liability for violations. Such liability both provides for redress for harms suffered by individuals and creates incentives for AI system developers and deployers to minimize the risk of those harms from occurring in the first place. As discussed below, in some cases, existing laws and regulations already provide baseline principles and frameworks to address harms caused by the use of AI. However, those laws are incomplete and inadequate in crucial respects.

---

[93] Artificial Intelligence Risk Management Framework, National Institute of Standards and Technology (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[94] Emanuel Moss, *et al.*, *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*, Data & Society (June 29, 2021), https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

Consider, for example, laws relating to civil rights and consumer protection. As the Federal Trade Commission (FTC), the Equal Employment Opportunity Commission (EEOC), the Consumer Financial Protection Bureau (CFPB), and the Department of Justice (DOJ) affirmed in their joint statement released on April 25, 2023, their enforcement authorities apply to discrimination and other harms arising from the use of automated systems.[95] The Department of Housing and Urban Development (HUD) also recently released a Renters Bill of Rights that recognizes the unlawfulness of algorithmic discrimination.[96] These agencies enforce several civil rights and consumer protection laws under which automated decision-making should be accountable:

- Section 5 of the FTC Act prohibits unfair or deceptive acts or practices. Under this authority, a practice – including one that is algorithm-driven – is unfair if it "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or competition."[97] The practice is deceptive if it involves a material "representation, omission or practice that is likely to mislead the consumer acting reasonably in the circumstances, to the consumer's detriment" because it is likely to affect the consumer's choice or conduct regarding the product.[98]

- The Dodd-Frank Act also prohibits unfair or deceptive practices and applies the same standard for unfairness and deception as the FTC Act does. In addition, the Dodd-Frank Act prohibits abusive practices, which are defined as (1) practices that materially interfere with a consumer's ability to understand a term or condition of a consumer financial product or service, or (2) practices that, with respect to the financial product or service, take unreasonable advantage of a consumer's lack of understanding of material risks, costs, or conditions; their inability to protect their interests; or their reasonable

[95] Rohit Chopra, *et al.*, *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems*, Federal Trade Commission, https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.
[96] *The White House Blueprint for a Renters Bill of Rights* (2023), https://www.whitehouse.gov/wp-content/uploads/2023/01/White-House-Blueprint-for-a-Renters-Bill-of-Rights.pdf.
[97] 15 U.S.C. § 45(n).
[98] Federal Trade Commission, Policy Statement on Deception (Oct. 14, 1983), https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf.

reliance on a covered person to act in their interests.[99]

- Title VII of the Civil Rights Act (Title VII), Title I of the Americans with Disabilities Act (ADA), and the Age Discrimination in Employment Act all prohibit several discriminatory employment practices that now often involve automated decision-making.[100] All of these laws prohibit employers and employment agencies from limiting, segregating, or classifying workers in any way that would adversely affect or tend to adversely affect their employment opportunities or employment status based on a protected characteristic. They all also prohibit discrimination in job advertising.[101] Title VII prohibits the use of ability tests to discriminate.[102] The ADA prohibits the use of selection criteria that screen out or tend to screen out disabled workers, and the administration of tests in a manner that reflects workers' disabilities instead of the factors the tests purport to measure.[103]

- The Fair Housing Act (FHA) prohibits several practices that adversely affect a person's opportunity to rent or buy a home based on their protected characteristics, and that are often enabled through automated decision-making. These prohibited practices include making the sale or rental of a dwelling unavailable on a discriminatory basis, or discriminating in the terms, conditions, or privileges of – or provision of services or facilities for – the sale or rental.[104] The FHA also makes it unlawful to cause an advertisement to be made or published that indicates a discriminatory preference regarding sale or rental of a dwelling.[105]

- The Equal Credit Opportunity Act (ECOA) prohibits creditors from discriminating in any aspect of a credit transaction.[106] Within thirty days of an adverse credit decision, ECOA requires the creditor to provide applicants with a written adverse action notice that

[99] *Unfair, Deceptive, or Abusive Acts or Practices,* Consumer Financial Protection Bureau (Mar. 2022) https://files.consumerfinance.gov/f/documents/cfpb_unfair-deceptive-abusive-acts-practices-udaaps_procedures.pdf.
[100] 42 U.S.C. § 2000e-2(a)(2); 42 U.S.C. § 12112(b)(1); 29 U.S.C. § 623(a)(2).
[101] 42 U.S.C. § 2000e-3(b); 29 C.F.R. § 1630.4(a)(1); 29 U.S.C. § 623(e).
[102] 42 U.S.C § 2000e-(2)(h).
[103] 42 U.S.C. § 12112(b)(6)-(7).
[104] 42 U.S.C. § 3604(a)-(b).
[105] 42 U.S.C. § 3604(c).
[106] 15 U.S.C. § 1691(a).

either (1) provides a statement of reasons for the adverse decision or (2) informs the applicant of their right to a statement of reasons and from whom the statement can be obtained.[107]

- The Fair Credit Reporting Act (FCRA) defines a consumer reporting agency as any entity that regularly engages in assembling or evaluating consumer credit data or other consumer data for the purpose of furnishing consumer reports to third parties, including through automated means.[108] FCRA restricts the furnishing of consumer reports, which are a consumer reporting agency's communications of any information bearing on a person's creditworthiness, credit standing or capacity, character, general reputation, personal characteristics, or mode of living, to be used in whole or in part for determining eligibility for personal credit or insurance or employment, or for other enumerated permissible purposes.[109]

Automated decision-making is also subject to a number of laws governing the public sector, where the use of AI requires careful management to protect the public without disrupting the administration of government programs. Title VI of the Civil Rights Act prohibits discrimination on the basis of race, color, or national origin under any program or activity that receives federal financial assistance, including through intentional exclusion from program participation or denial of program benefits, as well as criteria or methods of administration that appear neutral but have a discriminatory effect based on race, color, or national origin.[110] Title IX of the Education Amendments prohibits discrimination on the basis of sex in any federally funded education program. The Rehabilitation Act and the Americans with Disabilities Act prohibit discrimination against people with disabilities. Section 504 of the Rehabilitation act applies the same protections as Title VI to discrimination based on disability,[111] and Title I of the ADA extends this prohibition to any state or local government entity, regardless of the entity's receipt of federal financial assistance.[112]

[107] 15 U.S.C. § 1691(d)(2).
[108] 15 U.S.C. § 1681(f).
[109] 15 U.S.C. §§ 1681a(d), 1681b.
[110] 42 U.S.C. § 2000-d; 28 C.F.R § 42.104(b)(2).29 U.S.C. § 794(a).
[111] 29 U.S.C. § 794(a); 28 C.F.R § 42.503(b)(3).
[112] 42 U.S.C. § 12132.

All of these legal frameworks apply to AI-based systems, meaning regulators already have a toolbox to turn to when addressing certain harms stemming from AI-based systems. Nevertheless, in some respects, these laws are not fit for purpose and do not provide adequate accountability for AI harms. The greatest challenge in successfully enforcing a claim against AI harms under existing civil rights and consumer protection laws is that the entities developing and deploying AI are not always readily recognized as entities that traditionally have been covered under these laws. This ambiguity helps entities responsible for AI harms claim that existing laws do not apply to them.

For instance, companies that perform tenant screening for landlords or compile and furnish data for advertising have argued that they are not consumer reporting agencies under FCRA.[113] Under FCRA, if an adverse credit decision was based on data that would typically be included in a consumer report but was furnished by a third party other than a consumer reporting agency, the user of the data must disclose the nature of the data upon the affected consumer's written request, and must inform the consumer of their right to make this request when the adverse action notice is provided.[114] In other words, while a consumer would at least be told in an adverse action notice the key factors that hurt their credit score when data is provided by a consumer reporting agency, a consumer would have to affirmatively request in writing the nature of data provided by another third party that led to an adverse action. Even if a company is recognized as a consumer reporting agency, it is not liable for the adverse decision itself under FCRA.

Affected people also face a heavier burden with respect to discrimination claims arising from the use of AI. Generally, a plaintiff would first have to establish a prima facie case through direct evidence of an AI user's discriminatory intent, or by showing that they belong to a protected class, tried to pursue an opportunity for which they would be eligible, and received an adverse outcome while similarly situated people outside of that protected class did not experience the same negative outcome.

[113] *See, e.g.,* Ariel Nelson, *Broken Records Redux: How Errors by Criminal Background Check Companies Continue to Harm Consumers Seeking Jobs and Housing*, National Consumer Law Center at 29 (Dec. 2022), https://www.nclc.org/wp-content/uploads/2022/09/report-broken-records-redux.pdf.
[114] 15 U.S.C. § 1681m(b).

Due to the lack of transparency in AI uses, the plaintiff may not have the information needed to even establish a prima facie case. They may not even know whether or how an AI system was used in making a decision, let alone have the information about training data, how a system works, or what role it plays in order to offer direct evidence of the AI user's discriminatory intent or to discover what similarly situated people experienced due to the AI. Obtaining any of this information can require significant time, legal support, and technical expertise, putting affected people at a deep disadvantage when they may have to continue actively seeking other available opportunities while pursuing a legal claim. Agency enforcement is resource-intensive as well—HUD began investigating Meta's targeting of housing advertisements in 2018 in response to a complaint filed by the National Fair Housing Alliance, and these proceedings culminated in the DOJ's complaint and settlement with the company in 2022.[115]

To be sure, plaintiffs in cases under civil rights and consumer protection laws face obstacles and disadvantages even absent the use of automation. However, the use of AI can exacerbate or create new obstacles, and the laws may need to be changed or supplemented to account for these and provide for accountability.

Private sector use of automation in the moderation, generation, and recommendation of online content presents both a different existing legal regime and a distinct set of policy considerations. The most relevant aspects of existing law are Section 230 of the Communications Act and the First Amendment.[116] Section 230 generally shields providers of interactive computer services from federal and state civil claims, as well as prosecution under state criminal law, for claims that treat the provider as the publisher or speaker of user-generated content, including their decisions to "publish, edit, or withdraw" user-generated content from their services.[117] In CDT's view, this liability shield encompasses providers' use of automated content analysis tools as well as recommendation algorithms, as these are both

[115] *Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising*, Department of Justice (June 21, 2022) https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known.

[116] 47 U.S.C. § 230.

[117] *See Zeran v. America Online*, 129 F.3d 327, 332 (4th Cir. 1997). Section 230 has never shielded online services from prosecution under federal criminal law; also, (and especially relevant to questions arising from generative AI tools), Section 230's liability shield does not apply to intellectual property claims (§ 230(e)(2)).

features of a provider's approach to exercising editorial discretion over the publication of user's speech on their services.

This liability shield has been instrumental in establishing an open and vibrant online environment for speech, where intermediaries do not face legal risk for hosting, distributing, or otherwise facilitating access to huge volumes of user speech. It has also, by design, made it difficult for parties to successfully bring suit against online intermediaries for their use of automated systems for alleged amplification of messages e.g. from terrorist organizations.[118] This reflects an intentional choice by Congress, in enacting Section 230, to encourage online services to engage in content moderation without fear that doing so would increase their risk of being held liable for unlawful content that they mistakenly left online.[119] Given the strong protections for speech provided by the First Amendment, Congress could not legally compel online services to remove hate speech, disinformation, pornography, or other lawful speech from their services.

Section 230 is not an absolute immunity; for example, it has never prevented online service providers from being charged with violations of federal criminal law. Section 230 also does not shield online service providers from liability when the provider contributes to the "creation or development, in whole or in part" of the content at issue.[120] Courts, notably the Ninth Circuit in the *Roommates.com* decision, have held that Section 230 does not shield a provider who "contributes materially to the alleged illegality of the conduct."[121] So, for example, a provider who constructs a webform that requires a user to input the preferred race or gender of a potential roommate is not shielded by Section 230 from lawsuits alleging that such preferences are a violation of the Fair Housing Act.[122] For example, when a user/advertiser provides racially neutral housing ad content and targeting criteria to an ad-targeting system, and the ad-targeting system displays the ad to a racially disproportionate audience, that ad-targeting system may be

[118] *See generally, Gonzalez v. Google LLC*, 598 US __ (2023).
[119] *See* Jeff Kosseff, *Correcting the Record on Section 230's Legislative History*, Technology & Marketing Law Blog (Aug. 1, 2019), https://blog.ericgoldman.org/archives/2019/08/correcting-the-record-on-section-230s-legislative-history-guest-blog-post.htm.
[120] 47 U.S.C. § 230(f)(3).
[121] *See Fair Housing Coun., San Fernando v. Roommates.com*, 521 F.3d 1157, 1168 (9th Cir. 2008).
[122] *But see Fair Hous. Council v. Roommate.com, LLC*, 666 F.3d 1216 (9th Cir. 2011) (holding that the Fair Housing Act does not apply to the sharing of living units and so it is not unlawful for an online service provider to facilitate discrimination in the selecting of a roommate).

understood to be the cause of the racially discriminatory (and, under the Fair Housing Act, illegal) outcome, and it is unclear whether Section 230 applies.[123] Courts have also begun to examine the limits of Section 230 when considering products liability claims.[124]

The contours of Section 230 and whether interactive service providers should face greater accountability, particularly for use of automated content analysis tools as well as recommendation algorithms, remains an active debate. As discussed in Section II.C below, that debate already has begun with respect to generative AI. Section 230 does not necessarily shield the creators of generative AI tools from liability for the content that those tools produce, though both the speech created by generative AI tools and the editorial discretion exercised by online content hosts are nevertheless significantly shielded from restriction by government action by the First Amendment.

4. Government procurement reform

The public sector can help set an example across the digital ecosystem for greater AI accountability through its policies and rules for government procurement. As mentioned above, Title VI of the Civil Rights, Section 504 of the Rehabilitation Act, and Title I of the ADA all prohibit discrimination in the administration of government programs. Because government agencies are widely using AI for a range of functions, government procurement processes need reform to minimize harms from the use of AI.

One issue that reform efforts must address is the lack of a common definition of "artificial intelligence" for all agencies.[125] Without a common definition, agencies can decide unilaterally that an AI-driven system does not meet their chosen definition of AI and is therefore not subject

[123] *See* M. Ali et al., Discrimination through optimization: How Facebook's ad delivery system can lead to skewed outcomes, (Sep. 12, 2019) https://arxiv.org/pdf/1904.02095.pdf.
[124] Peter Karalis & Golriz Chrostowski, *Analysis: Product Claims Spike as SCOTUS Ponders Section 230 Fix*, Bloomberg Law (Mar. 2, 2023), https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-product-claims-spike-as-scotus-ponders-section-230-fix.
[125] For example, the National Artificial Intelligence Initiative Act of 2020, which focuses on AI research and development, provides a different definition for AI than the John S. McCain National Defense Authorization Act, whose definition is used in other AI legislation. The latter is also used in Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, which exempts AI embedded within common commercial products.

to obligations for responsible AI practices.[126] Therefore, there must be a whole-of-government definition for AI that all agencies must apply, and agencies must report the tools they acquire and how they are classified to confirm whether AI-driven systems are being classified as such.

Bringing accountability to government use of AI is made difficult by the limits to agencies' technical and workforce capacity, which causes agencies to rely more heavily on vendor assurances.[127] Without sufficient resources to thoroughly evaluate AI vendor proposals for the anticipated fitness for purpose, and for the potential risks of their systems, agencies are disincentivized to categorize AI-driven systems correctly and fulfill the obligations that come with acquiring AI, let alone to ensure responsible use on an ongoing basis.

NIST can push agencies to use their existing capacity to prevent AI harms by developing a standard for performance of algorithmic impact assessments. As addressed earlier, NIST's AI RMF is one of the few government sources of guidance for assessing an AI system's risks. In addition to possible use profiles for public sector uses of AI, NIST can establish a federal standard for impact assessments scoped to different levels of potential impact on civil rights and liberties, similar to NIST's creation of federal information processing standards for assessing whether risks of an information system have a low, moderate, or high security impact.[128] Agencies' decision-making regarding AI system acquisition should also be a public process guided by broad stakeholder input, especially from groups whose access to government services and benefits will be affected.[129]

---

[126] *See* NASA Office of Inspector General, NASA's Management of Its Artificial Intelligence Capabilities (May 3, 2023), https://oig.nasa.gov/docs/IG-23-012.pdf (describing how NASA has three different definitions for "artificial intelligence," and personnel choose the definition that aligns most with their own understanding of the term).
[127] Ross Wilkers, *Acquisition Shops Inside Government Need Workers Too*, Washington Technology (June 24, 2022), https://washingtontechnology.com/contracts/2022/06/acquisition-shops-inside-government-need-workers-too/368569/.
[128] National Institute of Standards and Technology, *Standards for Security Categorization of Federal Information and Information Systems*, FIPS Pub 199 (2004), https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf.
[129] Administrative Conference of the United States, Agency Use of Artificial Intelligence (Dec. 31, 2020), https://www.acus.gov/recommendation/agency-use-artificial-intelligence ("When an AI system narrows the discretion of agency personnel, or fixes or alters the legal rights and obligations of people subject to the agency's action, affected people or entities might also sue on the ground that the AI system is a legislative rule adopted in violation of the APA's requirement that legislative rules go through the notice-and-comment process.").

B. Accountability in practice: AI and Workers

In this section, we illustrate how the four accountability tools described above could be applied to the use of AI in the employment context, specifically the use of automated employment decision-making tools (AEDTs) and electronic surveillance and automated management (ESAM). In the sphere of hiring and employment, both anecdotal evidence and informal, unscientific surveys indicate that employers are increasingly deploying automated systems to surveil, manage, and make key employment decisions regarding their workers.[130] But due to a near-complete absence of transparency obligations surrounding these tools, the full prevalence and impact of these systems is largely obscure.

Because the use of AEDTs and ESAM in the hiring and employment context threatens great harm to workers' legal rights and livelihoods, accountability for their use, including close regulatory scrutiny aligned with the spirit of existing civil rights protections, is needed. Unfortunately, virtually no laws specifically govern those tools/practices, and the existing accountability frameworks are very limited in scope, outdated, and often ill-fit-to-purpose. There is scant transparency and only limited and ambiguous auditing requirements, and even those only really kick in when there is a suspicion that discrimination is occurring—something that is all but impossible in the absence of transparency. As described below, working with civil rights groups, CDT published in December 2022 the *Civil Rights Standards for 21st Century Employment Selection Procedures*, which describes auditing and transparency measures that would help provide greater accountability to the use of AI in the workplace.[131]

---

[130] *See, e.g.*, Dinah Wisenberg Brin, *Employers Embrace Artificial Intelligence for HR*, SHRM (Mar. 22, 2019), https://www.shrm.org/resourcesandtools/hr-topics/global-hr/pages/employers-embrace-artificial-intelligence-for-hr.aspx (2019 survey found that 83% of US employers use AI in HR); *see also*, Littler Mendelson, *The Littler Annual Employer Survey Report* (May 2022), https://www.littler.com/files/2022_littler_employer_survey_report.pdf (finding that 69% of surveyed employers use AI or data analytics in recruitment and hiring, while 29% use it for HR strategy and employee management).

[131] Matt Scherer & Ridhi Shetty, *Civil Rights Standards for 21st Century Employment Selection Procedures*, Center for Democracy & Technology (Dec. 5, 2022), https://cdt.org/insights/civil-rights-standards-for-21st-century-employment-selection-procedures/.

1. Automated employment decision-making tools (AEDTs)

With respect to employment discrimination laws, AI-powered hiring tools are subject to the same laws as traditional employee selection procedures, and employers can therefore be held liable for discrimination if an employer uses a tool that discriminates against a protected group. The Uniform Guidelines for Employee Selection Procedures (UGESPs), which are written into federal regulations, apply to the enforcement of Title VII of the Civil Rights Act (against private employers) and Executive Order 11246 (against federal contractors).[132] They state that employers should run tests to determine whether formal selection procedures have a disparate impact based on race, sex, national origin, color, or religion.[133] If they do, the UGESPs require an employer to *validate* the selection procedure to determine whether the selection procedure actually measures important job characteristics.[134] Unfortunately, the UGESPs have not been updated in nearly 50 years, and the procedures they describe are ill-suited for assessing validity or detecting potential sources of discrimination in sophisticated AI-powered tools. Further, the UGESPs do not apply to all protected classes – they pre-date the ADA and explicitly state that they do not apply to responsibilities under the Age Discrimination in Employment Act or the Rehabilitation Act.[135]

The lack of clear federal laws on auditing and disclosures has led to a lack of accountability in this space, with employers and vendors alike disclaiming responsibility for AI tools that are discriminatory or simply shoddy. On the one hand, employers rely on vendors' claims about validity and bias auditing of their automated employment decision tools. On the other hand, vendors—which have not been treated as employment agencies but increasingly are performing functions of employment agencies through their products—blame employers for relying so heavily on these tools.

Meanwhile, workers have too little insight to effectively scrutinize how they are being evaluated. Once they submit their resumes, they cannot observe how their resumes are reviewed, so they will not know whether they were discriminated against because of their

[132] 29 C.F.R. § 1607.2(A).
[133] 29 C.F.R. § 1607.4.
[134] 29 C.F.R. §§ 1607.5-1607.14.
[135] 29 C.F.R. § 1607.2(D).

names, zip codes, affiliations, or experiences. Without being told what data a gamified test, questionnaire, or video interview analysis will collect and analyze, and exactly why and how the selected assessment will analyze this data, workers cannot determine whether the assessment will evaluate them fairly, and disabled workers cannot determine whether they should seek reasonable accommodations under the ADA. This opacity makes it all but impossible for workers to compare decisions affecting them to those affecting similarly situated people, to show that an employer's use of these tools is in fact based on discriminatory intent, or to identify less discriminatory alternatives.

2. Electronic surveillance and automated management (ESAM) systems.

Artificial intelligence is also increasingly used to monitor and control workers' on-the-job activities using electronic surveillance and automated management (ESAM) systems. As in the context of hiring and other pre-employment decisions, federal laws like Title VII and the ADA prohibit employers from using AI in a way that discriminates against workers with protected characteristics. There are, however, no federal laws requiring employers to notify workers that they are being subjected to automated management systems or other forms of AI-driven monitoring in the workplace, much less explain how such systems work and how they are used in employment decisions.[136] Moreover, common law gives employers wide latitude to monitor and manage their workers, whether through technological or traditional means. As a result, there are few constraints on companies' ability to deploy ESAM in ways that threaten workers' health and safety, chill workers' right to organize, and exacerbate the power imbalance that already exists between employers and employees.

[136] On July 1, 2023, amendments to the California Consumer Privacy Act will take effect, making California the first state to require employers to inform workers about their workplace data collection practices. *See generally* Cal. Civ. Code § 1798.145(m)(4), available at https://www.caprivacy.org/cpra-text-with-ccpa-changes/#1798.145(m)(4) ("This subdivision [exempting employee-related data] shall become inoperative on January 1, 2023."). *See also, California Fails to Extend CCPA's Employee and Business-to-Business Data Exceptions*, Baker Botts (Sept. 12, 2022), https://www.bakerbotts.com/thought-leadership/publications/2022/september/california-fails-to-extend-ccpas-employee-and-business-to-business-data-exemptions. The law does not, however, address the use of artificial intelligence or other automated decision-making systems.

Addressing AI's role in employment discrimination requires a more modernized and proactive approach to accountability. For example, the *Civil Rights Standards for 21st Century Employment Selection Procedures* published by CDT describe auditing and transparency measures that would more effectively ensure that automated employment decision-making processes are designed and deployed to mitigate discriminatory outcomes and only evaluate workers' ability to perform essential job functions.[137] The *Civil Rights Standards* suggest requirements for both employers and vendors to prevent both parties from deflecting responsibility to each other for algorithmic discrimination.

These measures include pre-deployment auditing to identify and address potential discrimination risks against all protected characteristics, and annual audits thereafter to respond to unaddressed sources of bias and newly discovered discrimination risks. Employers and vendors would have to inventory the essential job functions for the positions for which their decision-making tools are used, and provide objective evidence that their decision-making tools directly measure workers' ability to perform essential job functions – not just correlation between decision-making criteria and job functions. This guidance document also recommends that employers and vendors be responsible for different types of notice tailored to the needs of different stakeholders: short-form disclosures for workers, audit summaries for regulators, and detailed recordkeeping for potential agency investigation. Further, the *Civil Rights Standards* prompt employers and vendors to examine how to better provide accommodations and alternative selection methods when the tools they use could potentially adversely affect a job candidate based on protected characteristics.

In addition to guiding private employers' acquisition of AEDTs and ESAM systems, the pre-deployment auditing and ongoing auditing requirements of the *Civil Rights Standards* can also improve accountability for how these tools affect workers in the public sector. Under the *Civil Rights Standards*, when a public sector employer looks to acquire an AEDT or ESAM system, their request for proposals (RFP) would require documentation of how the tool was designed and trained, how it was assessed for job-relatedness, bias risks, and privacy risks, and how bias and privacy risks were mitigated. When an agency puts out an RFP for a government contractor, the RFP would require prospective contractors to identify the AEDTs or ESAM systems they will

[137] *See* Matt Scherer & Ridhi Shetty, *Civil Rights Standards, supra* at n. 131.

use when carrying out the contract and provide documentation of how those tools are designed, trained, and assessed. As a result, agencies would share responsibility with their vendors who develop these tools, or with their contractors who deploy these tools, to fulfill the auditing obligations proposed by the *Civil Rights Standards*.

C. Accountability and Generative AI

In some respects, accountability for generative AI presents similar issues as other forms of AI. For example, the underlying models rely on training data, which may or may not be biased. Likewise, just as with other AI systems, it may not be clear why a generative AI application produced the outputs that it did. One core difference, however, is that unlike systems such as those that engage in automated decision-making, generative AI creates or produces various types of content, from text to images to video. Because speech and expression is the central output of generative AI, regulating that output in the form of imposing accountability necessarily implicates related human and legal rights related to speech and expression. For example, heavily restricting generative AI would harm free expression, creativity and innovation, and potentially run afoul of the First Amendment. Any regime for accountability for generative AI needs to recognize that reality.

The accountability landscape in the context of generative AI can be thought of in at least four phases:

- The creation of the foundation model (the underlying machine learning model such as GPT-3 or DALL-E upon which a variety of tools can be built);[138]
- The development of a specific tool, use-case, or deployment of a foundation model (e.g. a chatbot or a tool to help hiring managers summarize resumes);
- The deployment or use of those tools to generate content (e.g. by the user providing a prompt to a chatbot or by a software company using the tool to create code); and
- The further use or distribution of the outputs of those tools (e.g. a user posting the chatbot-produced text on social media, or the software company incorporating the AI-generated code in a consumer software product).

[138] *See* Rishi Bommasani, *et al*., *On the Opportunities and Risks of Foundation Models* (Jul. 12, 2022), https://arxiv.org/pdf/2108.07258.pdf.

In each of these phases, different actors make specific choices that shape the output and operation of a model or the tools it supports, which in turn may increase or decrease the risk of harm to individuals. Understanding these phases, and the choices and actors involved in each, is helpful for mapping the potential opportunities for accountability.

In the creation and development phases, the company, research group, or other developer of a generative AI model makes a number of choices that affect the ultimate operation of the system and its potential effects on individuals' rights.[139] This includes the choice of data used to train the foundation model; as in other machine learning contexts, the underlying training data will shape the capabilities and potential biases of the model or tool and its output. Generative AI tools may generate text that includes personal or confidential information that was part of the training data set. Training data will also likely include copyrighted material. In addition to questions around liability for scraping data to use for training purposes, generative AI tools raise distinct questions about the extent of potential copyright violations involved when the tool produces text or an image that closely mimics an artist or author's distinctive style.[140]

In addition to the initial training of the model, generative AI tools are often fine-tuned for particular purposes and to achieve particular goals through "reinforcement learning with human feedback".[141] This step in the development of the tool involves providing additional training data and employing human feedback as part of the process to train the tool towards providing (or not providing) different types of responses. (This could include, for example, fine-tuning the tool to not provide personal information of non-public figures when prompted by a user.[142])

The goals and design interface choices selected by the developer of a generative AI tool can have a significant impact on how users interact with the tool and how its outputs are received.

---

[139] *See* Laura Weidinger, *et al.*, *Taxonomy of Risks Posed by Language Models* (June 2022), https://dl.acm.org/doi/10.1145/3531146.3533088.
[140] *See* Stephen Wolfson, *The Complex World of Style, Copyright, and Generative AI* (Mar. 23, 2023) https://creativecommons.org/2023/03/23/the-complex-world-of-style-copyright-and-generative-ai/.
[141] Paul Christiano, *et al.*, *Deep Reinforcement Learning from Human Preferences* (Feb. 2023) https://arxiv.org/abs/1706.03741
[142] Katie Malone, *What Do Chatbots Know About Us, and Who Are They Sharing It With?*, Engadget (Apr. 7, 2023), https://www.engadget.com/what-do-ai-chatbots-know-about-us-and-who-are-they-sharing-it-with-140013949.html.

This includes the crucial decision of whether to make the model or tool available in open- or closed-source formats, which will affect how much anyone beyond the original developer of the tool will be able to affect the design and operation of it, including adding or removing safeguards against abuse. Developers of generative AI tools set the basic boundaries for how it may be used, including whether to provide users the ability to change certain parameters about the tool (as OpenAI does in its chatGPT "playground") to allow for more experimentation and flexibility in the use of the tool. Developers also make choices about whether a tool can interact with or be integrated into other tools, e.g. providing an API that enables users to connect a chatbot to a social media account to enable automatic posting, or using a chatbot to generate responses to search queries.

Ultimately, developers make decisions about whether a particular tool is designed to support humans' creative or expressive endeavors, such as the Midjourney Bot, Adobe's generative model Firefly that is integrated into Photoshop, or a chatbot's ability to produce stories, news articles, and essays, or whether it is intended for more instrumental outcomes, such as answering a user's specific questions, creating computer code, or assisting an employer in making hiring decisions. While chatbot tools today such as ChatGPT and Bard are broadly capable of all three of those functions, the "generic chatbot" interface is by no means a necessary component of generative AI tools,[143] and may confound user expectations about the reliability and utility of the text these tools produce.[144]

In addition to the developer, the user of the tool can have a significant impact on its output. Users may, for example, reveal personal information about themselves or others in the prompts they provide to a chat bot, which may then become incorporated into the tool's corpus of text and potentially revealed to other users in the future. Users may also (intentionally or not) prompt tools to generate illegal content, such as a defamatory (i.e. false and reputation-damaging) statement about a real individual, or images that violate obscenity, privacy, intellectual property, or rights-of-publicity laws.

[143] Michal Luria, *Your ChatGPT Relationship Status Shouldn't Be Complicated*, Wired (Apr. 11, 2023), https://www.wired.com/story/chatgpt-social-roles-psychology/.
[144] Larry Neumeister, *Lawyers Blame ChatGPT for Tricking Them into Citing Bogus Past Cases in Court*, LA Times (June 9, 2023), https://www.latimes.com/world-nation/story/2023-06-09/chatgpt-lawyers-cite-bogus-case-law.

Finally, many of the risks to individuals from generative AI outputs depend on what the user of the tool proceeds to do with that information. Some of these risks occur directly between the user providing the prompt and the tool. For example, a user who asks questions about high-stakes topics such as medical issues, mental health concerns, public safety, or financial management, and then relies on authoritative-sounding but incorrect answers could put themself at risk of physical or financial harm. Other risks stemming from generative AI content will be felt primarily if that content is distributed more broadly by being published to the web or spread via email or text messaging. The decision to distribute AI-generated content to a broader audience may be made unilaterally by the user (copying the content and posting it on another service) or it may be facilitated by the developer (by providing an API to integrate a tool with an account capable of mass communication).

Other risks will accrue when the individual decides to use or act on the information provided by the tool in a way that affects others. For example, a software company that buys and uses a generative AI tool for producing code and incorporates that code into a consumer software product without adequate procedures for reviewing and testing the code may cause harm to its customers (e.g., if the AI-generated code had a security flaw).

As this discussion shows, accountability in the context of generative AI is complex: it implicates multiple actors at different phases, each of which may affect the risk of harm, and all of this is in the context of expression and speech. Notwithstanding these complications, each of the four tools described above can play a role in providing accountability.

Transparency. In some respects, the same considerations around transparency apply equally to generative AI, such as the need to disclose the training data used and explain how the system arrives at its outputs. Users should also know if they are interacting with a generative AI system. In the case of a chatbot, that may be relatively evident. But because some generative AI tools are designed to mimic human interaction, it may not always be so clear. A user, for example, should know whether they are messaging with an actual customer service representative or just seeing AI-generated outputs.

Another aspect of transparency of particular importance in the context of generative AI is making clear when text, video, or images are generated by AI as opposed to a human.

Generative AI tools are—by design—remarkably good at producing content that appears to be generated by a human. As discussed above, that reality increases the risks associated with deep fakes, fraud, and widespread digital influence operations.[145]

Partially in response to these risks, companies and people have begun building systems designed to detect whether content is created by a generative AI system. However, these are currently largely ineffective.[146] But developers of generative AI systems can facilitate the detection of AI-generated content by enabling their software to embed "watermarks." One possible approach to watermarking text, for instance, modifies the pattern of text generation enough to allow detection in only a very short sample of text, without affecting the quality of the generated text.[147] Some developers of AI image tools have announced plans to embed watermarks.[148] If developers of generative AI systems were to commit to watermarking their outputs (in perhaps a standardized way, to further ease detection), it would be easier for users to know when they are seeing synthetic content. Watermarking is not a perfect solution; even if all major AI companies were to watermark their outputs, users hoping to evade detection (perhaps in order to deceive a target audience) might turn to open-source generative AI systems configured to not watermark outputs. Users could also de-watermark text by, for example, passing watermarked outputs through another piece of software that paraphrases the text.[149] The tug-of-war between watermarking systems and users who might try to evade the watermarks is an area of active research. But watermarking may nonetheless provide a benefit, allowing detection of a significant amount of AI-generated content that a user might encounter online.

---

[145] *See* Josh A. Goldstein, *et al.*, *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (Jan. 2023), https://arxiv.org/pdf/2301.04246.pdf.

[146] Armin Alimardani & Emma Jane, *We Pitted ChatGPT Against Tools for Detecting AI-Written Text, and the Results are Troubling*, The Conversation (Feb. 19, 2023), https://theconversation.com/we-pitted-chatgpt-against-tools-for-detecting-ai-written-text-and-the-results-are-troubling-199774. .

[147] *See* John Kirchenbauer, *et al.*, *A Watermark for Large Language Models* (Jun. 6, 2023), https://arxiv.org/pdf/2301.10226.pdf.

[148] See, e.g., Kyle Wiggers, *Microsoft Pledges to Watermark AI-Generated Images and Videos*, TechCrunch (May 23, 2023), https://techcrunch.com/2023/05/23/microsoft-pledges-to-watermark-ai-generated-images-and-videos/.

[149] *See* Kalpesh Krishna, *et al.*, *Paraphrasing Evades Detectors of AI-Generated Text, But Retrieval is an Effective Defense* (Mar. 23, 2023) https://arxiv.org/pdf/2303.13408.pdf.

Audits and assessments. In many respects, audits and assessments should play a similar role in providing accountability for generative AI systems as they do for other forms of AI. The same questions around who should conduct such audits and against what standards apply here as well. One form of assessment of particular relevance to generative AI is the use of red-teaming. The goal of red-teaming is to create prompts that cause the model to generate text that is likely to cause harm or otherwise undesirable behaviors. It enables testing of the efficacy of safety measures and identifies weaknesses that can then be mitigated.

Laws and liability. As discussed above, a key law governing liability for online content is Section 230. Courts have yet to address the application of Section 230 to generative AI scenarios. When they do, key considerations will be to what extent the content created by the generative AI system can be considered "user-generated content," and to what extent the generative AI system will have participated in the "creation or development" of the content in question.[150] If a tool is effectively providing snippets or recombining existing user-generated content in its training data and providing those statements in response to a user prompt, courts may find that the tool is operating no differently from a search engine, which also provides existing user-generated content in response to prompts, and the tool generating those outputs should thus be shielded by Section 230.[151] But, given chatbots' propensity for "hallucinations," or the generation of false statements that do not appear in their training data, it appears that existing generative AI tools already go beyond merely recombining existing user-generated content and are creating, or substantially contributing to, novel content. Courts likely will find that, at least in some applications, a tool that generates novel content is not shielded from legal claims by Section 230.[152] These decisions will have significant implications for the development, deployment, and use of generative AI tools.

Given the multiple actors and decision points involved in the creation, development, use, and distribution of generative AI tools and the content they produce, there is likely no one-size-fits-all model for liability for generative AI. The user who prompts a tool to draft a convincing news article that falsely claims that a private individual is a child molestor, and then

[150] 47 U.S.C. § 230(f)(3) (definition of "information content provider").
[151] *See* Jess Miers, *Yes, Section 230 Should Protect ChatGPT and Other Generative AI Tools*, TechDirt (Mar. 17, 2023), https://www.techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/.
[152] *See* Matt Perrault, *Section 230 Won't Protect ChatGPT*, LawFare (Feb. 23, 2023), https://www.lawfareblog.com/section-230-wont-protect-chatgpt.

distributes that article widely, is the primary culpable party for the negative effects of that publication of defamation. The user who poisons himself after following a chatbot's suggested home remedy for curing a cold is himself the injured party, and will want to hold the creator of the chatbot accountable for his injuries.

Rather than create a new liability regime out of whole cloth, at least initially courts will likely apply existing areas of law (such as privacy and data protection, tort law, copyright, and criminal law) to the harms produced by generative AI. For example, people who use generative AI to perpetrate scams could be prosecuted for fraud, extortion, or harassment; face investigation by the Federal Trade Commission for unfair and deceptive trade practices; or face civil litigation for claims such as fraud, intentional infliction of emotional distress, harassment, defamation and intellectual property violations.

Applying existing laws to generative AI will raise key questions about how core elements of these areas of law apply: Do the creators of foundation models or developers of generative AI tools owe a duty of care to users to prevent harms? If so, for what kinds of harms? Can mens rea standards of "knowing" or "willful" action apply to the operation of generative AI tools? How should liability be apportioned across the creators of foundation models, the developers of specific tools, and the users of those tools who provide the prompts or make use of their outputs? In what circumstances should "aiding and abetting" or principles of secondary liability apply? What effects will different approaches to liability have on the further development of generative AI models and tools?

It is necessary to begin grappling with these questions across different generative AI use-cases. There is a risk that courts could interpret existing law in a way that leaves individuals who have suffered injuries as the result of generative AI tools with no source for recovery, for example by finding that neither the user who provided a prompt nor the developer of a chatbot tool had the requisite knowledge and intent necessary to be legally responsible for a defamatory statement generated by the chatbot. However, it is important to remember that not every adverse or problematic outcome from a generative AI tool violates a law, and thus there may not be any existing remedy for that harm under law.

Government procurement. As with other AI systems, to the extent the government purchases generative AI tools or services, government procurement rules can influence the development of standards for generative AI. The government may also have an important role to play in providing the resources needed to develop foundational models or other elements that can be used to develop generative AI systems. Such models require enormous resources to develop and operate and as a result may only be in the reach of a small number of large, well-resourced companies. Whether through grants, procurement, or otherwise, the government could help facilitate the development of more widely available models and, in so doing, require accountability measures that in turn serve as best practices for others to follow.

As policymakers, courts, industry, researchers, and civil society consider accountability frameworks for generative AI, CDT suggests accountability frameworks should incorporate the following principles to help guide those inquiries:

- Recognize the **different roles** that different actors play in the creation, development, use, and distribution of generative AI tools and outputs.
- Emphasize the importance of **risk assessment and mitigation**.
- Focus expectations/obligations for different actors based on the **degree and type of control** they have over the system and the harms it may generate.
- Encourage creators of foundation models and developers of tools to prioritize making their models and processes **transparent and explainable**.
- Identify **risk mitigation measures** that can be incorporated into the **fine-tuning and design decisions** for a given tool.
- Encourage developers of a tool to **monitor use of the tool for abuse** and continue to update the tool and the safeguards around its use.
- Assign **primary liability for illegal content or conduct to the proximate actor** whose actions led to the harm (e.g. the user who decided to distribute a defamatory statement or the software company who incorporates AI-generated code without appropriate testing and review).
- Remind users of generative AI to **be cautious about the veracity, reliability, and legality of the information** they create with generative AI tools.

Conclusion and Recommendations

Accountability requires the involvement of all stakeholders: the developers, deployers, and users of AI, government and regulators, civil society, communities and individuals who may be harmed or otherwise affected by the use of AI, and more. We focus here on some of the steps NTIA and the Executive Branch more generally should take to increase AI accountability with respect to each of the four accountability tools discussed above, as well as the role of Congress.

Transparency and Explainability. The government should take steps that set an expectation of transparency around the development, deployment, and use of AI. In higher-risk settings, such as where algorithmic decision-making determines access to economic opportunity, that may include transparency requirements. The government should also consider whether to facilitate the standardization of transparency formats (e.g., model and system cards) to help users and others more easily understand the information provided about AI systems. That could involve multistakeholder convenings run by NTIA or having NIST or another body develop voluntary standards for certain forms of transparency. The Executive Branch could also support through NSF grants or otherwise the development of techniques such as watermarking to increase transparency around when content has been generated by an AI tool.

Audits and Assessments. Audits and assessments are critical to accountability, but fundamental questions about how to conduct those effectively remain unanswered. For example, auditors need standards to apply, and those standards embody important value judgments. The Executive Branch can help facilitate the development of meaningful audit standards. For example, NIST has announced its intent to build on the AI Risk Management Framework by developing "profiles" to help different developers, deployers, and end users of AI apply the Risk Management Framework to their contextual setting (e.g., a user profile for the use of AI in employment decisions). These profiles could help inform the development of auditing standards and help shape accountability in practice. Given the potential value of the NIST profiles to shape public behavior, NTIA should encourage NIST to ensure that profiles are developed with robust participation by civil society and independent experts, and that profiles meaningfully address the elements of trustworthy AI identified in the Risk Management Framework and the Blueprint for an AI Bill of Rights.

Law and Liability. As explained above, federal agencies have recognized that their existing enforcement authorities apply to AI-based discrimination and other harms. In the short term, each agency with authority to enforce civil rights or consumer protection laws should conduct investigations into entities who engage in harmful AI practices. Each agency should also publish guidance that:

- Provides illustrative examples of how covered entities may contribute to harms that violate the laws that the agency is authorized to enforce when using common types of AI-based systems; and
- Offers criteria that covered entities should look for or questions they should ask that would help ensure compliance with laws relevant to the agency's authority when selecting an AI-based system to deploy.

Beyond this guidance, each agency should also (to the extent its existing authorities allow) pursue new rules to establish standards for covered entities to, among other things,

- Provide meaningful notice regarding the functions and risks of the systems they use to those who interact with these systems;
- Conduct algorithmic impact assessments of their systems, document assessment results, and publish summaries of assessment results; and
- Clarify how definitions of covered entities may apply to third-party vendors.

Government Procurement. The government should adapt its procurement policies and regulations to better ensure accountability when it procures AI. That includes identifying a definition for "artificial intelligence" that ensures stronger agency reporting on AI procurement. More broadly, agencies will need to implement the forthcoming OMB guidance on use of AI by the government,[153] which will create an opportunity for agencies to take steps to ensure that their procurement policies align with the principles identified in the AI Bill of Rights.

[153] *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety* (May 4, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

Congress also can take concrete steps to increase transparency and accountability in the design, development and use of AI tools, including through appropriations provisions, oversight of relevant federal agencies, and steps such as hearings, convenings, and/or the creation of a Commission to highlight best practices and novel innovations to address potential harms.

Congress is also developing and considering a wide array of proposed laws that seek to increase the accountability of AI systems such as the Algorithmic Accountability Act. NTIA and the Administration more broadly should support that continued legislative development. An essential threshold step to protect against AI-related harms is to pass comprehensive privacy legislation that affirmatively limits data collection, processing, and transfer in the United States. Because data is a key input for training AI and algorithmic systems, enacting data protections can go a long way in protecting against AI-related harms. CDT has supported the American Data Privacy and Protection Act (ADPPA) in 2022, which would impose strong data minimization requirements on companies that would largely limit collection of sensitive data to that which is strictly necessary to provide a service requested by the individual.

ADPPA also has a strong civil rights section that would prevent discrimination based on protected classes in data practices. That section also requires AI and algorithmic transparency. Large companies would have to create algorithmic impact assessments for algorithms that create consequential risks of harm that describe, among other things, a description of the design process and methodologies, the purpose of the algorithm, a description of the data used to train the algorithm, and a description of the steps the company has taken to address several types of harm. That section also requires a pre-deployment algorithmic design evaluation on any company that develops an algorithm for broad use, in furtherance of a consequential decision, that requires the evaluation of the design, structure, and inputs of the algorithm and an analysis of how the company reduced the risks identified above.