Making AI more explainable to protect the public from individual and community harms

Dr. Nicol Turner Lee
Senior Fellow, and Director of the Center for Technology Innovation, Brookings Institution

Written Statement to the U.S. Senate AI Insight Forum on Transparency, Explainability, Intellectual Property, & Copyright

November 29, 2023

Majority Leader Schumer, Senators Rounds, Heinrich, and Young, and other distinguished Members of the Senate, I thank you for the invitation to participate in one of many dialogues about the current and future implications of artificial intelligence (AI), especially today's conversation focused on the transparency and explainability of models and the systems, as well as the impacts on intellectual property and copyright protections. The ability of everyday people to understand when and how AI systems are used and how they affect their eligibility for a range of activities—from credit worthiness, mortgage approvals, college admissions, to employment—is of critical importance to deconstruct the opaque technology. As AI becomes more ubiquitous, consumer engagement is one of many critical aspects of transparency. Those who design, deploy, and license AI must also provide more visible explanations and disclosures of what consumers should expect from the technology, as well as assurances around its safety and trustworthiness. Ordinary people are increasingly the subject of AI, and do not have agency over it either because the use of these tools is not disclosed, or the massive collection and surveillance of individual and community data by companies are not governed.

Given this framing, I will touch on the following areas in my written statement:

- The importance of AI consumer disclosures as fundamental tenets to systems transparency and explainability, and the application of easily accessible tools and processes that strengthen consumer trust and confidence.
- The ongoing accountability and oversight of AI technologies that ensure the civil and human protections of users and protect individuals and communities from the extraction of their ideas and intellectual property.
- The lack of redress for consumers and communities when harmed by AI, generative systems, and future technologies whose impacts often result in reputational risks and other vulnerabilities.

These points will lead me to conclude my statement with three recommendations that I distinguish by urgency. The *first* is for Congress to act now on universally available and applied consumer disclosure standards that can be applied to high-risk, autonomous decision-making systems to stave off considerable harms, especially those that are discriminatory. The *second* is for Congress to explore the creation of a formal commission that includes various aspects of government, private industries, academic disciplines, and civil society to develop recommendations around both self-regulatory guidance and more prescriptive legislation that increase consumer trust and assurance, while interrogating the agility and function of existing civil rights and other regulatory regimes to protect more vulnerable consumers. Finally, the *third* recommendation is that there should be some mechanism to protect individuals and communities who have been evidently harmed, exploited, or coopted by AI decisions that offers recourse from developers, deployers, or licensees to address their reputational and economic losses.

I. What exactly are transparent and explainable AI systems?

Experts continuously emphasize the difficulty in understanding why an AI model behaves as it does, given that what is ultimately created is iteratively shaped by real-world applications and contextual frames. That is why there is an increasing need and call for openness, which is underscored by the fact that systems can be deceived or fail on tasks, even those easily performed by humans.[i] Because transparency in AI is a multifaceted concept, addressing technical cadences, along with broader socio-technical considerations are only parts of solutions to address the complexity of models and the systems where the internal workings of AI systems, or the "black boxes", are invisible to users.

On the positive side, when AI is applied to more general consumer contexts, it can serve as a digital helper, and support the completion of various tasks, including booking an airline ticket, or engaging government services. But concerns about online biases, consumer privacy, and the overreach of AI persist, which make calls for more transparency both timely and relevant. For example, certain financial services products, including "Buy Now Pay Later" programs target certain consumers and spending digital profiles with a lack of regulatory guidance, including the protections of existing consumer credit laws.[ii] Additionally, other research related to AI and financial services indicate that variables, such as whether a consumer uses a mac or PC are considered in assessing credit worthiness, and thereby, pose ethical questions and legal implications related to potential biases.[iii] Further, in mortgage approvals, the integration of AI in credit risk models has unveiled accuracy gaps between different demographic groups. A recent study assessing alternative credit-scoring models unveiled a significant 5 to 10% accuracy gap in predictive tools between lower-income families and minority borrowers when compared to higher-income, non-minority counterparts.[iv] In financial services alone, these inaccuracies emphasize the need for ongoing evaluation and potential adaptations in lending practices for fairness and equity, as well as greater awareness by consumers of AI practices, especially the norms of qualifications.[v]

II. The importance of consumer disclosures to avert discrimination and other malfeasance

Online biases can and have led to a set of intended and unintended consequences, particularly for federally protected groups. The AI decision making ecosystem is equally vulnerable to the proliferation of online synthetic media, or "deepfakes" that can spread mis- and disinformation and generate realistic, harmful content without subjects' consent. Any national governance framework must prioritize the strengthening of public trust in AI systems, while encouraging those who develop, deploy, and license these models to act more responsibly and openly.

Most recently, some Members of Congress have introduced the use of digital watermarks to denote the provenance of media (whether authentic or synthetic), and even the White House's recent Executive Order on AI stated plans for standardizing rules for these watermarks.[vi] However, watermarks alone will be insufficient, and can be leveraged by bad actors to potentially worsen the problem. Researchers have been able to subvert all the watermarking methods currently available and even insert false watermarks to create false positives.[vii] Overall, some experts do think that it could be part of an arsenal of tools to mitigate the risks of synthetic media in the future, with the caveat that this will be one of many components to raise consumer awareness.

Consumer trust and confidence in AI systems could also benefit from existing multistakeholder processes, including an Energy Star-inspired ratings system that reveals the accuracy and efficiency of such models, as well as the flaws and other shortcomings. In my own research, I argue that an Energy Star-inspired ratings system that is currently guiding the multi-stakeholder process in the consumer appliances market could be applicable to

AI by collaborating on a core set of operational standards, fostering a culture of transparency through some type of open labeling of AI systems, and potentially revealing when systems fall short. Such a rating system could also incentivize data scientists and developers to take ownership of their creations, conduct thorough testing, and incorporate appropriate auditing tools.[viii]

The United States has a series of multi-stakeholder processes that lend themselves to self-regulatory guidance around norms and standards. The creation of easily accessible and easy to understand consumer disclosure requirements by companies developing, deploying, and licensing AI and other emerging technologies should not be complicated, but instead urgent to encourage safe, and trustworthy systems, especially in high-risk domains.

III. Transparency in civil and human rights

AI is already integrated into countless aspects of daily life, and making recommendations about medical treatment, hiring and retention, and credit scores.[ix] In the criminal justice context alone, algorithmic decision-making systems have been used to deploy police officers, conduct face recognition surveillance, and determine criminal sentences.[x] In these and other contexts, the decisions that AI tools make can have serious consequences on the exercise and preservation of existing civil and human rights, which heighten the need for greater transparency to ensure its lawful compliance.

To date, the White House's Blueprint for an AI Bill of Rights has articulated, in principle, the need for algorithmic discrimination protections via impact assessments and plain language reporting for new AI systems.[xi] Most recently, a bipartisan group of congressional representatives proposed the *National AI Commission Act*, which would be responsible for regulating AI.[xii] Senators Wyden, Booker, and House Member Clarke have introduced the Algorithmic Accountability Act, which would require assessments of algorithms used for the "high impact" decisions.[xiii] While these actions are prudent and necessary, there are some areas where more proactive, civil rights governance, along with some reasonable expectation of legal compliance with laws that protect equal access and opportunities for consumers, are necessary. With a common understanding that biases are inherently baked into AI systems – whether it be the norms and values of the developer the quality of the training data, or the context for deployment, demonstrated differential treatment or disparate impact of these technologies must be considered, and disclosed when inadequacies or potential threats arise, like, for example, the complexion-related detection flaws in facial recognition technologies (FRT) that are widely used by law enforcement and are at greater risk of defendant misidentification[xiv]. Like other high-impact AI implementations, FRT and other technologies require clear standards and guidelines, as well as a sea change in transparency requirements that are made known to consumers.

We also need to know about the potential for bias and discrimination that algorithmic systems have *before* they disproportionately affect people's lives by denying them credit or a home loan or subjecting them to public arrest due to fundamental flaws in the technologies. The absence of transparency throughout the life cycle of the AI models can foreclose on the civil and human rights of users, especially through explicit discrimination and inferential assumptions baked into the norms of automated decision-making tools.

IV. The lack of transparency encroaches on creative rights

In recent months, we are also seeing disproportionate impacts and fears of AI on whole communities. Earlier this year, over 170,000 Hollywood writers and actors went on strike in support of better working conditions, including protections from AI used to replicate or replace their work. It was the longest strike in the history of both the Writers Guild of America and SAG-AFTRA.[xv] These creatives ultimately achieved an outcome, in large part, because they had an outcry of public focus and support.

The concerns of the writers, actors, and creators mor broadly were legitimate because I, too, worry if my forthcoming book will be summarized without attribution after three long years of gathering and writing original content. For more vulnerable populations, particularly those who come from diverse backgrounds or meager circumstances where ideas lend themselves to greater opportunities for economic stability, how does the lack of transparency encroach on their intellectual property - which often emanate from their unique lived experiences? A recent report from the World Economic Forum predicted higher job churn than usual over the next four years in two job areas: (1) supply chain and transportation, and (2) media, entertainment and sport.[xvi] When generative AI can compete with humans to create outputs that are just as good in half the time, we need to worry about an oversaturation of the markets, one that will disproportionately harm diverse voices. In a recent Brookings blog, my co-author, Regina Ta, and I also revealed that in the case of generative AI, it is only training on a limited set of languages, which means that the more accurate and robust narratives coming from diverse communities are being omitted.[xvii]

Even if AI were trained on a completely diverse and representative dataset, concerns around how the data was collected and used still permeate, and whether there was full license or respect of copyright to do so. For example, as a curated technology that relies on the digital artifacts, generative AI has been trained on datasets comprised of many thousands of books without their authors' knowledge or consent.[xviii] While several lawsuits already have been filed on this front, again, who speaks for communities without unfettered access to legal resources, or a general awareness of the cooptation of their ideas, especially in more sophisticated extractions of voice and image. Unfortunately, the increasing ubiquity of generative AI, and other AI formats make it more difficult for consumers to simply opt out without experiencing the costs of such exclusions.

V. Where is the social or financial compensation for unknown AI harms?

Recently, Porcha Woodruff became the sixth known person to be falsely accused of a crime by facial recognition technology in Detroit, Michigan. She was pregnant when she was arrested in front of her family and had to post a $100,000 personal bond to be released from jail.[xix] She is now suing the city. A couple of years ago, an algorithmic system used for determining care for chronically ill patients consistently excluded Black patients because the model relied upon the input variable of how much people paid into hospitalization.[xx] AI approval systems have systematically denied mortgages to Black applicants, and AI hiring tools have been known to discriminate against female and minority applicants.[xxi] These are just a sample of the many consequential examples of harms.

Biased AI systems have cost people their jobs, damaged their health, and eroded reputations. However, those on the receiving end of this discrimination often lack a clear path toward social or financial redress for the harm that AI systems caused. There is no explicit private right of action that lets individuals protect themselves from algorithmic or other forms of AI discrimination. The absence of a national data privacy standard that is fundamental to limiting the amount of data collected on individuals and their communities contributes to the exploitation of everyday consumers. But it's unrealistic to think that consumers will take their grievances to abstract AI models. People discriminate, not computers. How we address and resolve the social and economic costs to consumers is important because it also encourages developers, deployers, and licensees of AI to act more responsibly and minimized their possible indemnification from fallible and possibly unlawful systems and their results.

VI. Policy recommendations

The ideas offered in my written statement prompt a more serious dialogue that connects the implications of AI transparency and explainability with a need for more carefully deliberated universal standards on AI use, along with increasing attention to the nimbleness of existing regulatory regimes focused on protecting civil rights and intellectual property, as well as copyright protections. In the interest of time, I will share three proposals with Congress that could prompt immediate and longer-term actions.

1. **We need universal AI standards and easily accessible consumer disclosures to operationalize transparency.** Concepts of AI explainability and transparency are not just about redefining the technical cadence of such models. There are social costs and consequences, and ones that can imbue a series of personal and community risks. As mentioned, the "black box" nature of algorithmic systems is a significant obstacle toward expanding public trust, and it often includes not only the inner workings of the algorithm itself but the nature of its implementation by its deployer. Congress should consider easily accessible and readable consumer disclosures on high-risk and high-impact models so that consumers are aware of their use and can either make choices or interrogate the bases for decision-making. But reasonable standards and expectations of an AI tool must be clearly explicated by companies and known by consumers to make these disclosures helpful. More important, having federal privacy standards will be foundational to prescribe permissible and non-allowable forms of data, including those that ascertain one's federally protected characteristics, and trigger the potential for either intentional discrimination or disparate impact treatment.
2. **Congress should act on the creation of a formal commission to develop adaptable and easily interpretable regulatory regimes and governance strategies to establish the safe and trustworthy AI design and deployment that include civil and human rights protections.** As a technology with the potential to affect every facet of society, the regulation of AI could be sector specific, but must be connected to the overarching guiding principles that protect civil and human rights. A dedicated commission could be well suited to coordinate this implementation of principles and is referenced in the recent draft legislation by legislators. But membership must be multi-faceted with a focus on generating clear proposals for consideration by Congress, and plausible calls to protect the public interest from AI harms.
3. **The creation of a reasonable private right of action for individuals who have been harmed by AI could protect consumers from social and economic costs of intended and unintended consequences, while incentivizing better industry practices.** An explicit private right of action would empower individuals to defend themselves from AI harms in high-risk domains, and further incentivize developers to investigate and mitigate bias in their products. Right now, there exists no reasonable and accessible framework for individuals to know when the AI has treated them comparatively different from similarly situated people and objects. There is also no recourse for the often embarrassing and outright damaging circumstances when the technology forecloses on civil and human rights, or intellectual property. Having a conversation around how both sides can benefit from core technologies is critical and returns agency back to consumers who have become passively consumed into AI's rapid development. Such a recommendation can be nationally directed or embedded within sectors who work to create more equitable and fair systems. In the end, impacted populations should not have to feel limited recourse to AI systems that make mistakes or discount their visibility. They should also not be burdened financially or socially for wrongful decisions that have far reaching consequences. How such a framework could be operationalized could be a product of the formal commission produced above or explored by agencies with distinct jurisdiction over high-risk areas in partnership with civil society, civil rights, and consumer action organizations interested in preserving public interest goals.

[i] Sara Brown, April 20, 2021. "Machine Learning, Explained." MIT Press, Machine learning, explained | MIT Sloan.

[ii] Vivek Astvansh and Chandan Kumar Behera, October 15, 2023. The Hidden Risks of Buy Now, Pay Later. The Conversation, The hidden risks of buy now, pay later: What shoppers need to know (theconversation.com)

[iii] Aaron Klein, April 11, 2019. Credit denial in the age of AI. Brookings, Credit denial in the age of AI | Brookings

[iv] Edmund Andrews, April 6, 2021. How Flawed Data Aggravates Inequality in Credit. Stanford University, How Flawed Data Aggravates Inequality in Credit (stanford.edu)

[v] Ibid.

[vi] "Public Trust in AI Technology Declines amid Release of Consumer AI Tools." MITRE, September 19, 2023. Public Trust in AI Technology Declines amid Release of Consumer AI Tools (mitre.org)

[vii] Makena Kelly, "Watermarks aren't the silver bullet for AI misinformation" The Verge, October 31, 2023. Watermarks aren't the silver bullet for AI misinformation | The Verge

[viii] Nicol Turner Lee, 'Mitigating Algorithmic Biases through Incentive-Based Rating Systems', in Justin B. Bullock, and others (eds), The Oxford Handbook of AI Governance (online edn, Oxford Academic, 14 Feb. 2022), Mitigating Algorithmic Biases through Incentive-Based Rating Systems

[ix] Kanadpriya Basu et al. "Artificial Intelligence: How is It Changing Medical Sciences and Its Future?." Indian journal of dermatology vol. 65,5 (2020): 365-370. Artificial Intelligence: How is It Changing Medical Sciences and Its Future?

[x] Molly Callahan, Algorithms Were Supposed to Reduce Bias in Criminal Justice—Do They?, February 23, 2023. The Brink, Algorithms Were Supposed to Reduce Bias in Criminal Justice—Do They? (bu.edu)

[xi] The White House, "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People," October 2022

[xii] Office of Congressman Ted Lieu, REPS LIEU, BUCK, ESHOO AND SEN SCHATZ INTRODUCE BIPARTISAN, BICAMERAL BILL TO CREATE A NATIONAL COMMISSION ON ARTIFICIAL INTELLIGENCE, June 20, 2023. REPS LIEU, BUCK, ESHOO AND SEN SCHATZ INTRODUCE BIPARTISAN, BICAMERAL BILL TO CREATE A NATIONAL COMMISSION ON ARTIFICIAL INTELLIGENCE (lieu.house.gov)

[xiii] "Booker, Wyden, Clarke Introduce Bicameral Bill to Regulate Use of Artificial Intelligence to Make Critical Decisions like Housing, Employment and Education," Office of Corey Booker, September 21, 2023.

[xiv] Dawn Zapata, "New Study finds AI-enabled anti-Black bias in recruiting," Reuters, June 18, 2021; Rachel Goodman, "Why Amazon's Automated Hiring Tool Discriminated Against Women," ACLU, October 12, 2018; Will Douglas Heaven, "Predictive policing algorithms are racist. They need to be dismantled." MIT Technology Review, July 17, 2020; Kashmir Hill, Eight Months Pregnant and Arrested After False Facial Recognition Match, August 6, 2023. Eight Months Pregnant and Arrested After False Facial Recognition Match | The New York Times.

[xv] Gene Maddaus, "WGA Votes to Ratify Contract, Officially Ending One of Hollywood's Longest Strikes," Variety, October 9, 2023; Gene Maddaus, "SAG-AFTRA Approves Deal to End Historic Strike," Variety, November 8, 2023

[xvi] World Economic Forum, Future of Jobs Report 2023, May 2023, Future of Jobs Report 2023 (weforum.org)

[xvii] Regina Ta and Nicol Turner Lee, "How language gaps constrain generative AI development," Brookings, October 24, 2023.

[xviii] Alex Reisner, Revealed: The Authors Whose Pirated Books are Powering Generative AI, August 19, 2023. Revealed: The Authors Whose Pirated Books are Powering Generative AI | The Atlantic

[xix] Kashmir Hill, Eight Months Pregnant and Arrested After False Facial Recognition Match, August 6, 2023. Eight Months Pregnant and Arrested After False Facial Recognition Match | The New York Times

[xx] Tom Simonite, A Health Care Algorithm Offered Less Care to Black Patients, October 24, 2019. , A Health Care Algorithm Offered Less Care to Black Patients | Wired

[xxi] Emmanuel Martinez and Lauren Kirchner, The secret bias hidden in mortgage-approval algorithms, August 25, 2021. The secret bias hidden in mortgage-approval algorithms | AP News; Rachel Goodman, Why Amazon's Automated Hiring Tool Discriminated Against Women, October 12, 2018. Why Amazon's Automated Hiring Tool Discriminated Against Women (aclu.org)