# Written Statement for the U.S. Senate AI Insight Forum: Transparency, Explainability, Intellectual Property, & Copyright

*November 21, 2023*

As generative AI systems grow in capability, AI is poised to increase in importance and relevance far beyond today's usage[1]. Our growing use of a small handful of online platforms powered by increasingly more capable AI systems has created broad consensus around the world for transparency. Policymakers abroad[2], leading AI companies at home[3], industry experts[4], the Biden-Harris Administration[5], and members of this very chamber[6] all agree that AI systems need external parties to verify their safety.

However, despite AI's profound direct influence on American life, and broad consensus on the need for external access, direct access[7] to leading AI systems by individuals who are not directly compensated by the organizations they study is virtually impossible in today's research ecosystem — because providing untrusted parties with access to sensitive internal AI systems is logistically difficult, and poses security vulnerabilities. There is a fierce debate around exactly how transparency between AIs and disinterested parties should be logistically accomplished. The debate centers around a fundamental tradeoff between use and mis-use of relevant information in the transparency process. On one hand, external researchers want access to data

[1] Brandon, John. "What Spending 12 Billion Hours Per Day On Social Media Has Taught Us." *Forbes*, 12 Mar. 2023, https://www.forbes.com/sites/johnbbrandon/2023/03/12/what-spending-12-billion-hours-per-day-on-social-media-has-taught-us/?sh=5d8cd6406e13.

[2] The European Union passed the Digital Services Act which includes articles that require Very Large Online Platforms and Very Large Online Search Engines to subject themselves to scrutiny by external parties.

[3] The Biden-Harris Administration secured voluntary commitments from 15 leading AI companies to perform internal and external security testing of their AI systems before their release.

[4] In a recent survey, 98% of industry experts indicated that AI labs should employ third party audits, pre-deployment risk assessments, dangerous capabilities evaluations, and red teaming over their AI systems. Garfinkel, Ben et al. "Towards Best Practices in AGI Safety and Governance." *Center for the Governance of AI*, 17 May 2023, https://cdn.governance.ai/AGI_Safety_Governance_Practices_GovAIReport.pdf.

[5] The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence includes language around consulting with third-party evaluators to develop AI model evaluation tools and testbeds.

[6] The Kids Online Safety Act includes a transparency section that empowers independent, third-party auditors to inspect covered platforms. 47 Senators currently sponsor this act.The Platform Accountability and Transparency Act provides secure pathways for independent research. 6 Senators currently sponsor this act. The Bipartisan Framework for U.S. AI Act includes a section on transparency which would provide independent researchers access to data necessary to evaluate A.I. model performance.

[7] (i.e., can execute arbitrary projects against an AI's code, weights, data, and/or user logs)

and AI systems to perform research — and platforms want to facilitate the public confidence in their systems that external researchers can provide. On the other hand, AI platforms have legitimate security, legal, trade secrets, privacy, competitiveness, trust & safety, public relations, and financial cost concerns with facilitating access to their secure, proprietary AI systems running over sensitive user data — especially from untrusted parties. This begs the question, what level of access for the purpose of AI transparency and explainability justifies the security, legal, trade secrets, privacy, competitiveness, trust & safety, and public relations risks and costs of such access?

In our view, this tension — between access and not — has a clear resolution. Mis-use centers NOT on whether external researchers can answer appropriate research questions but rather on the miscellaneous data and trade secrets they might observe in the process, otherwise known as collateral information leakage. To conceptualize this notion, consider public venues where a bag search is required to uncover whether an individual is in possession of illegal drugs. In some cases, this requires venue security to physically search ALL contents in each bag, an expensive, time-consuming, privacy-invading procedure. In other cases, venues enlist the services of drug-sniffing dogs, which are trained to detect concealed illegal drugs, increasing the privacy, effectiveness, and efficiency of narcotics surveillance. In this scenario, there is almost no detectable tradeoff between privacy and security.[8] It is a non-issue as a result of technological solutions that filter through irrelevant, invasive information to only detect the rare bits of desired information - are you in possession of drugs?

The same is possible for external AI research. Using modern privacy-enhancing technologies (PETs), it is newly possible (from a technical perspective) for an external researcher to ask and answer virtually any statistical question about an internal AI system in a way that prevents other questions from inadvertently being answered. This moves the thorny tradeoff between business and civil society rights to a simpler question - should this specific question about this specific AI system be answered? While there are challenging edge cases[9], by and large, an external researcher can ask and answer almost any statistical question about

[8] Trask, Andrew. "Safe Crime Detection - Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance." *i am trask*, 5 Jun. 2017, https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/.
[9] For example if a platform has very few users, it can be difficult for differential privacy to effectively mask information about specific individuals while facilitating study about the group. However, naturally, the most influential AI systems have immense numbers of users.

an internal AI system and its impacts on groups of users without ever directly seeing the AI system or the data it runs on — and this answer can be verified using robust cryptographic verification systems.[10]

For example, consider the question of whether an AI system correlates with mental health problems among the American people. Given that Americans spend roughly 14% of their time awake interacting online[11], this is an important question about the quality of the American way of life. And notice, the question does not pertain to the AI system itself — but to the immense group of people who use it. The goal of safety research is to reveal or speculate about impacts on the lives of large groups — not proprietary intellectual property or data about specific individuals (except in the most extreme cases of individual user harm). It is — in part — this focus on overall trends in people that sets the stage for transparency optimism.

Yet today, using legacy transparency approaches, this question about AI's effect on mental health is virtually impossible to answer. It requires access to data about an AI's behavior — and data about the health of an AI's users. Data about an AI's behavior with various users is sensitive data that exists within AI platforms — data they're not going to let leave their secure facilities (because of legitimate concerns around privacy, security, trade secrets, legal, competitiveness, etc.). On the other hand, data about the health of an AI's users exists in medical institutions, such as the Centers for Disease Control and Prevention (CDC),[12]. Naturally, this data is highly sensitive and highly regulated. Using legacy transparency techniques, this mental health question demands that these two types of datasets — located in (at least) two locations from which they cannot be removed — must be moved into the same computer system for joint study. This is the dilemma that blocks AI transparency.

This dilemma means, a researcher traveling to the headquarters of an AI platform cannot surface this type of insight. A researcher using a static API built by an AI platform cannot surface this type of insight. A researcher working exclusively with open data cannot surface this

[10] We encourage the readers of this document to consider any question — and we would be happy to describe the infrastructure necessary to facilitate it while mitigating these concerns.

[11] 2 hrs and 14 mins * 331.9 million ppl. Buchholz, Katharina. "Where People Spend the Most & Least Time on Social Media." *Statista Daily Data*, 26 Apr. 2022, www.statista.com/chart/18983/time-spent-on-social-media/.

[12] "Public Health Data Systems that Provide Mental Health Information." *National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health*, 28 Apr. 2023, https://www.cdc.gov/mentalhealth/data_publications/index.htm.

kind of insight — because neither dataset can be safely released to the public. And for these reasons, the closest research insights we have to understanding the question of AI and mental health describe AI impacts at a high level — such as MIT's high-level correlation between universities adopting AI-powered platforms and their students developing mental health issues[13]. No study truly performs joint analysis between specific algorithms and the before-and-after mental health state of users because legacy transparency infrastructure — onsite access, API/web-app access, open data, etc. — cannot safely facilitate this type of research.

However, this question is no longer blocked by technical infeasibility because it is now possible to perform joint analysis on two datasets — from two locations — without ever seeing the datasets and without ever requiring them to be moved to the same location. An external researcher could — in theory — sit in their pajamas in a New York apartment, propose a computation to be run across a secure dataset at the CDC in Atlanta and a secure dataset at an AI platform in Silicon Valley — and generate the answer to a question like this. And in doing so, they would never learn anything about any specific medical patient. They would never learn the proprietary secrets of an AI platform. They would never need to enter a secure building. Data from the CDC would never need to leave the custody of the CDC. And data from the AI platform would never need to leave the control of the AI platform. Using a combination of techniques such as secure enclaves, secure multi-party computation (SMPC), homomorphic encryption, differential privacy, and zero-knowledge proofs — it is possible for an external researcher to answer this question — and learn absolutely no other private information about these datasets — or require the owners of these datasets to disclose anything else to anyone else.

This has a profound implication on transparency, namely related to legacy privacy, security, trade secrets, legal, competitiveness, trust & safety, and public relations risks. Instead of asking, "Should this researcher be trusted with access to sensitive information?" which is really about asking, "Do we trust this person to not mis-use our data?" — the transparency question is entirely about, "Should this question be answered?" and "Is this the right way to ask the question?". The profoundness of this shift in terms of facilitating AI transparency cannot be overstated. It is the difference between today — with limited and restricted external access by

[13] Walsh, Dylan. "Study: Social media use linked to decline in mental health." *MIT Sloan School of Management*, 14 Sep. 2022, https://mitsloan.mit.edu/ideas-made-to-matter/study-social-media-use-linked-to-decline-mental-health.

external researchers and no access to private third-party data — and tomorrow — where Americans only use AI systems that are verified and known to be safe.

Yet, achieving this vision of tomorrow still requires work. While research for this technology has reached an important threshold of maturity, and projects such as the Christchurch Call Initiative on Algorithmic Outcomes have proven the viability[14] of privacy-enhancing technology's applications for increased AI transparency and explainability, it is not yet widely deployed. OpenMined has worked with pioneering partners, including Microsoft's LinkedIn, Dailymotion, and X/Twitter, to deploy such transparency infrastructure for their production recommender systems and enable external civil society researchers to conduct research on such systems; but more industry participation is needed. Furthermore, minimally viable versions of this technology need to be made robust, third-party data (such as that at the CDC) needs to be made available through it, and regulatory channels must be constructed to politically legitimize this novel, poorly understood infrastructure into its proper place in society.

But, if the AI community is successful in doing so, perhaps AI can achieve perfect transparency around important safety questions while protecting user privacy and mitigating legitimate business concerns. Perhaps the 14% of American life spent guided by AI can be verified to be safe, secure, and trustworthy. Perhaps the future of AI can advance equity and civil rights, stand up for consumers and workers, promote innovation and competition, and advance American leadership around the world because we had the foresight to install the appropriate transparency infrastructure, ensuring that governments, regulators, civil society, and AI product developers have the proper eyes and ears to follow AI's progress and ensure its fruitfulness. In OpenMined's view, AI can achieve all these things and more; we welcome Congress's role in steering our collective AI future towards one with an effective transparency regime and look forward to partaking in this formative discussion on how we get there.

[14] "2023 Leaders' Summit Joint Statement" *Christchurch Call*, 11 Nov. 2023, https://www.christchurchcall.com/assets/Documents/Christchurch-Call-Leaders-Summit-2023-Joint-Statement-ENG.pdf.