July 25, 2023

Response to the Australian Government's Department of Industry, Science and Resources Request for Consultation on Responsible Artificial Intelligence

To the Australian Government,

Thank you for opening up this request for consultation and for taking the time to review this response. I am writing on behalf of OpenMined, a global not-for-profit community of 16,000+ members with significant experience facilitating external access to proprietary artificial intelligence (AI) systems. For the past 6 years, we have worked as a global community on freely available open-source software for this purpose. We are not auditors, nor are we policy analysts. We are technologists who focus on a specific aspect of the AI assurance problem: what is the best way to *facilitate external access to internal systems*?

We are delighted by all the actions the Australian Government has already taken to responsibly advance the development and use of AI, including releasing and piloting an AI Ethics Framework that is consistent with the OECD's Principles on AI, issuing guidance on public sector adoption of AI as part of the Australian Government Architecture, and investing in a National AI Centre to uplift responsible AI practices, among other contributions. Our response will focus on the ways in which the Australian Government can build upon these initiatives and direct future actions toward the development and piloting of AI assurance mechanisms as a central pillar of any responsible AI program.

Discussions around requiring assurance mechanisms like independent audits and external assessments of AI systems have been widespread and recurring in recent responsible AI conversations. We seek to provide an operational framework that leverages modern software approaches to execute such requirements. We also seek to share some of the significant technical advancements that have made this possible, in that it may be useful as you continue to be a global leader in responsible AI.

About Us

OpenMined is a global non-profit that specializes in the problem of facilitating access to internal, secure data and AI systems by external parties — often using privacy-enhancing technologies (PETs). We began in 2017 as an informal, online meeting point for researchers from different fields relevant to this problem. Since then, our free online courses have garnered 20,000+ student registrations, our freely available, open-source software has been contributed to by over 500 open-source contributors, and numerous academics, startups, industry, and policy professionals have launched or pivoted their careers during their time in our community. Broadly speaking, we are well positioned to represent the perspective of the PETs community in this testimony to you — and its relevance to your continued support of responsible AI practices.

You can find our work in the White House's 2023 report on Privacy-Preserving Data Sharing and Analytics[1], the Royal Society's 2023 report on PETs, and in the United Nations' 2023 PETs guide. We have worked with the private sector, building and deploying privacy technology with: NVIDIA, Google (1, 2, 3 privacy tech partnerships), Meta (1, 2, 3, 4, 5 privacy tech partnerships), Microsoft (1 public), and Twitter (1, 2, privacy tech partnerships). We also create freely available and public technical tutorials on some of the use cases for applying PETs (e.g., How to audit an AI model owned by someone else).

Our work with the public sector includes co-launching and chairing the United Nations Privacy Enhancing Technology Lab, deploying on the United Nations Global Platform, and testing and improving our software in collaboration with UN PET Lab Members —which include the United States as represented by the Census Bureau's emerging technology group, xD. Our work in civil society includes a number of academic institutions. Taken together, we have worked extensively on this problem with public, private, and civil society partners across the world.

As a non-profit, we are supported by a diverse pool of funders. Our top two funders are Georgetown University and the Alfred P. Sloan Foundation, followed by Meta, the Office of the Prime Minister of New Zealand, and additional funders. As a result of this generous support, *everything we produce we give away for free,* and — amidst a sea of startups — we are one of the only well-funded non-profits focusing on the end-to-end problem of external access to internal systems. We believe we are uniquely positioned to offer neutral, objective, and informed comments on the procedures by which the Australian Government may mitigate any potential risks of AI and support safe and responsible AI practices through transparency mechanisms like audits.

[1] Search "PySyft"

Most Relevant Work

Starting as early as 2018, OpenMined began to narrow its general focus on external access to the specific challenge of AI assurances in the form of algorithmic oversight. This culminated in a partnership with Twitter to advance algorithmic transparency, which was announced publicly in early 2022. Over the following year, we worked closely with Twitter's Machine Learning Ethics, Transparency, and Accountability (META) team on the challenge of facilitating external research on their production recommender algorithm. The project focused on measuring political bias during the 2020 U.S. Presidential Election. While we had partnered with online platforms on privacy technology many times before, facilitating the study of *large-scale production* systems running on *private* user data comes with unique privacy, security, and intellectual property (IP) challenges. Finding solutions required work across our own engineers and researchers in partnership with Twitter's product, engineering, research, legal, and policy teams. Even with buy-in from senior stakeholders — who launched the project voluntarily — external access to internal algorithms is not a trivial task.

By the end of 2022, the project was successfully expanded into a new scope: the Christchurch Call Initiative on Algorithmic Outcomes (CCIAO). New Zealand's Prime Minister at the time, Jacinda Ardern, and France's President, Emmanuel Macron, launched the CCIAO as a partnership between OpenMined, the United States White House, New Zealand, Microsoft, and Twitter to support the creation of new technology to better understand the impacts that algorithms and other processes may have on terrorist and violent extremist content.

In order to study those impacts, CCIAO partners had to overcome challenges around how to protect user privacy and proprietary information, how to investigate impacts holistically across society, and how to achieve reproducibility, affordability, and scale for independent researchers. As the first project under this initiative aiming to address these challenges, OpenMined has developed privacy-enhancing software infrastructure to enable an outside researcher to perform research on an internal production algorithm. This software is called PySyft.

PySyft allows model owners to load information relating to production AI algorithms into a server, where an external researcher can then send a research question without ever seeing the information in that server. In doing so, privacy, security, and IP need not block external accountability of algorithms. External researchers can extract answers to important questions without ever obtaining direct access to the data driving those answers — or the spaces (physical buildings) in which that data is housed.

During Phase 1 of CCIAO, the partners have worked together to deploy PySyft infrastructure in a way that does not infringe on the data-use policies of partner online service providers and still makes it possible for an external researcher to download results from private data within these secure sites. In line with our charitable mission, as supported by our donors, all of the infrastructure we built in this effort is freely available under open-source licenses, and participation by all parties has been on a voluntary basis. Based on this experience, we share some of what we've learned below.

Framing the Problem: Insufficient Access to AI Systems Hides Harms

Many policymakers agree that independent evaluation is required to demonstrate the safety and effectiveness of an AI system. Despite this agreement, often, independent evaluation or other oversight exercises, like audits, are blocked. To audit an organization's AI system, an auditor needs access to internal, proprietary computers, data, models, APIs, and talent at that organization — surfacing valid concerns about privacy, security, intellectual property, cost management, multi-stakeholder coordination, and result verification.

This *access problem* underpins a second problem: employees who work at an AI organization often cannot see the downstream impacts of their AI models because the records of such impacts are at *other organizations*. Most AI organizations do not have ground-truth demographic data about their users (i.e., income level, race, religion, etc.) with which to study whether their algorithm appears to favor or disfavor one group or affect user welfare. In fact, most AI organizations have little information, if they even have any information at all, about broader user welfare.

For example, an engineer who works on a movie recommender system for a streaming company may not know that the recommender reduces a user's sleep. Furthermore, while a sleep app might have this type of information, the sleep app would (and should) be reluctant to share this type of information with the movie streaming company because of privacy, security, IP, and competitive dynamics. The inability of the streaming company to answer the specific question, “Is my AI recommender causing sleep problems?” is an access problem created by a lack of access to the necessary information (i.e., user sleep patterns) required to answer that question. Crucially, the streaming company doesn’t need or want user data on their sleeping patterns - they only want the answer to their specific question, which is hidden in 3rd party data they cannot access for very valid reasons.

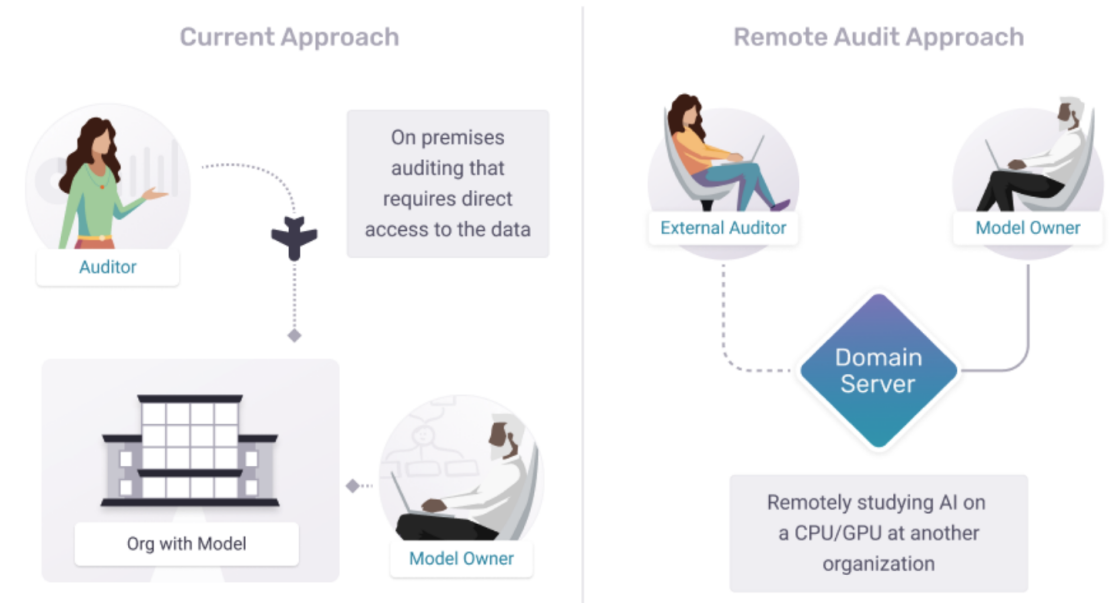
When these two access problems - accessing an internal system and accessing 3rd party data - are unsolved, it is possible for an AI to cause harm without being detected.

Solutions to the AI Access Problems:

1. **Remote Audits**

To get closer to a world in which we can mitigate the risks of AI and support safe and responsible AI practices, we describe a new way forward that alleviates these access problems — remote audits. Remote audits offer a query-based system with one key capability: an external auditor can propose a question about an AI system to its owner and related third parties, and — if they approve the question — the auditor will be able to download the answer to that question without the auditor, AI owner, or third parties learning anything beyond what the group explicitly approves.

To do this, an AI owner copies their AI system into a server, which exposes an API to an external auditor. Meanwhile, third parties load their data into their own servers, making them available for use in AI audits. The auditor can then download fake/mock pieces of the AI system and fake/mock third-party datasets which imitate the AI owner and 3rd parties' real assets and use those fake pieces to create audit software. The auditor then sends this proposed audit software to the AI owner and third parties for approval. If all actors consent (through manual review or an automatic process), they allow the code that was *developed using fake/mock data* to be run *on the real data*, creating the answer to a specific question about an AI system (e.g., does my AI model hurt users' sleep). Taken together, the external auditor can download the answer to a specific question without any parties learning any other information in the process (e.g., without the AI owner seeing the sleep data or vice versa). This facilitates more precisely scoped collaboration between internal and external parties in keeping with the concept of [structured transparency].



Remote audits strongly curtail the limitations of current external access approaches — which struggle to balance the tradeoffs between auditor access and AI owner risk. Approaches that prioritize the AI owner's protection over the auditor's access might require limiting audits to high-trust individuals, constraining auditors' field of inquiry to a subset of shareable data, constraining auditors' field of inquiry to a narrow pre-built API, or running an auditor's query for them — restricting the ability of an auditor to create and verify their own results. On the other hand, other approaches can grant an auditor too much freedom with an AI system, exposing owners to extraneous privacy, security, and IP risks that exceed the audit's intended scope. Furthermore, none of the current access approaches properly leverage third-party data about user welfare to reveal the downstream impacts of AI products on real people after they leave the AI system — which is the true motivation behind many audits. Auditors frequently rely on heuristics about the quality of AI systems — test datasets, simulated environments, etc. — while the true impact on users' lives remains unverified.

2. Phased and Continuous Audits

We also propose that any assurance mechanism (e.g., remote audits) should be conducted on a phased and continuous basis. Taking inspiration from other sectors, AI auditing is comparable to phases of a drug trial. Early in a drug trial, scientists test a drug in simulated environments — in test tubes in a lab. Late in a drug trial, scientists measure whether people were helped or harmed when they took the drug in the course of their normal life. Similarly, early in an AI evaluation, scientists test an AI model in simulated environments — in video-game worlds or against test datasets. Later, scientists can run an AI model in the real world — and create an aggregate measurement of whether people's lives get better or worse when they use it.

AI auditing is like a drug trial because neither the audit nor the drug trial is actually about the model or drug. Audits and drug trials are about people — and whether their lives get better or worse when they use new technology in the real world. Early in a drug trial, the test tube is meant to *predict* or *approximate* whether the drug would be likely to cause harm — but even if the drug looks promising — nobody *really* knows at that point. The test tube is not conclusive. It's just early evidence. Early in an AI audit, the test dataset or video game environment is the same. The test dataset is meant to *predict* or *approximate* whether an AI model would be likely to cause harm — but nobody *really* knows for sure based on a test dataset or simulation. A test dataset is not conclusive. It's just early evidence.

The ultimate test — for both novel medications and for AI models — is deployment in the real world. Pick any AI algorithm in your mind, perhaps a: self-driving car algorithm, newsfeed algorithm, content recommender, advertising algorithm, job recommender, credit score, or GPT. Ask yourself a challenging question: how would the owner of an AI algorithm be able to know whether its model makes people's lives better or worse? To test whether an AI model is safe, fair, unbiased, non-toxic, responsible, etc. — you must figure out whether people's lives get better or worse when they use it. This analogy can be useful for drawing an instructive model of risk management from the pharmaceutical sector in that it offers a useful oversight framework to ensure AI systems are safe and remain safe over time in various contexts.

I will finish by saying governments who want to mitigate any potential risks of AI while also supporting safe and responsible AI practices must address the access problems — accessing an internal system and accessing 3rd party data — or else AI will be permitted to cause harm that is difficult to detect and prevent. Phased and continuous remote audits are just one assurance mechanism that is available and mature enough to begin to address the access problems in a meaningful and effective way, but we would be happy to discuss others with your office if of interest.

Our hope through these comments is that the Australian Government will consider the utility of assurance mechanisms when determining the next steps for supporting their responsible AI ecosystem, and we look forward to Australia's continued leadership and progress in this space.

Sincerely,

Lacey Strahm
Policy Lead, OpenMined