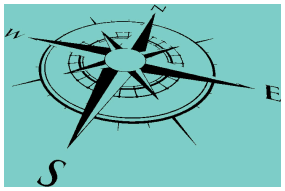
Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

Contents



Trustworthy AI Workgroup (TAIWG)

Stakeholder(s):

Contributors :
Listed authors are those who contributed substantive ideas and/or work to this report. Contributions include writing, research, and/or review for one or more sections; some authors also contributed content via participation in an April 2019 workshop and/or via ongoing discussions. As such, with the exception of the primary/corresponding authors, inclusion as author does not imply endorsement of all aspects of the report.

Corresponding Authors :
†Miles Brundage (miles@openai.com), Shahar Avin (sa478@cam.ac.uk), Jasmine Wang (jasminewang76@gmail.com), Haydn Belfield (hb492@cam.ac.uk), and Gretchen Krueger (gretchen@openai.com) contributed equally and are corresponding authors. Other authors are listed roughly in order of contribution.

Miles Brundage :
Corresponding Author - OpenAI

Shahar Avin :
Corresponding Author - Leverhulme Centre for the Future of Intelligence; Centre for the Study of Existential Risk

Jasmine Wang :
Corresponding Author - 4Mila; 29McGill University; Work conducted in part while at OpenAI.

Haydn Belfield :
Corresponding Author - Leverhulme Centre for the Future of Intelligence; Centre for the Study of Existential Risk

Gretchen Krueger :
Corresponding Author - OpenAI

Gillian Hadfield :
OpenAI; University of Toronto; Schwartz Reisman Institute for Technology and Society

Heidy Khlaaf :
Adelard

Jingying Yang :
Partnership on AI

Helen Toner :
Center for Security and Emerging Technology

Ruth Fong :
University of Oxford

Tegan Maharaj :
Mila; Montreal Polytechnic

Pang Wei Koh :
Stanford University

Sara Hooker :
Google Brain

Jade Leung :
Future of Humanity Institute

Andrew Trask :
University of Oxford

Emma Bluemke :
University of Oxford

Jonathan Lebensold :
Mila; McGill University

Cullen O'Keefe :
OpenAI

Mark Koren :
Stanford Centre for AI Safety

Théo Ryffel :
École Normale Supérieure (Paris)

JB Rubinovitz :
Remedy.AI

Tamay Besiroglu :
University of Cambridge

Federica Carugati :
Center for Advanced Study in the Behavioral Sciences

Jack Clark :
OpenAI

Peter Eckersley :
Partnership on AI

Sarah de Haas :
Google Research

Maritza Johnson :
Google Research

— continued next page

Stakeholders (continued)

Ben Laurie :
Google Research

Alex Ingerman :
Google Research

Igor Krawczuk :
École Polytechnique Fédérale de Lausanne

Amanda Askell :
OpenAI

Rosario Cammarota :
Intel

Andrew Lohn :
RAND Corporation

David Krueger :
Mila; University of Montreal

Charlotte Stix :
Eindhoven University of Technology

Peter Henderson :
Stanford University

Logan Graham :
University of Oxford

Carina Prunkl :
Future of Humanity Institute

Bianca Martin :
OpenAI

Elizabeth Seger :
University of Cambridge

Noa Zilberman :
University of Oxford

Seán Ó hÉigeartaigh :
Leverhulme Centre for the Future of Intelligence; Centre for the Study of Existential Risk

Frens Kroeger :
Coventry University

Girish Sastry :
OpenAI

Rebecca Kagan :
Center for Security and Emerging Technology

Adrian Weller :
University of Cambridge; Alan Turing Institute

Brian Tse :
Future of Humanity Institute; Partnership on AI

Elizabeth Barnes :
OpenAI

Allan Dafoe :
Future of Humanity Institute; University of Oxford

Paul Scharre :
Center for a New American Security

Ariel Herbert-Voss :
OpenAI

Martijn Rasser :
Center for a New American Security

Shagun Sodhani :
Mila; University of Montreal

Carrick Flynn :
Center for Security and Emerging Technology

Thomas Krendl Gilbert :
University of California, Berkeley

Lisa Dyer :
Partnership on AI

Saif Khan :
Center for Security and Emerging Technology

Yoshua Bengio :
Mila; University of Montreal

Markus Anderljung :
Future of Humanity Institute

Workshop Participants :
Acknowledgements — We are extremely grateful to participants in the April 2019 workshop that catalyzed this report, as well as the following individuals who provided valuable input on earlier versions of this document ... None of these people necessarily endorses the content of the report.

David Lansky

Tonii Leach

Shin Shin Hua

Chris Olah

Alexa Hagerty

Madeleine Clare Elish

Larissa Schiavo

Heather Roff

Rumman Chowdhury

Ludwig Schubert

Joshua Achiam

Chip Huyen

— continued next page

Stakeholders (continued)

Xiaowei Huang

Rohin Shah

Genevieve Fried

Paul Christiano

Sean McGregor

Tom Arnold

Jess Whittlestone

Irene Solaiman

Ashley Pilipiszyn

Catherine Olsson

Bharath Ramsundar

Brandon Perry

Justin Wang

Max Daniel

Ryan Lowe

Rebecca Crootof

Umang Bhatt

Ben Garfinkel

Claire Leibowicz

Ryan Khurana

Connor Leahy

Chris Berner

Daniela Amodei

Erol Can Akbaba

William Isaac

Iason Gabriel

Laura Weidinger

Thomas Dietterich

Olexa Bilaniuk

Miles Brundage :
attendees of a seminar talk given by author Miles Brundage on this topic at the Center for Human-Compatible AI (CHAI).

Vision

A more trustworthy AI development ecosystem

Mission

To identify mechanisms for supporting verifiable claims for AI development

Values

Artificial Intelligence: Artificial intelligence has the potential to transform society in ways both beneficial and harmful.

Trust: Beneficial applications are more likely to be realized, and risks more likely to be avoided, if AI developers earn rather than assume the trust of society and of one another. This report has fleshed out one way of earning such trust, namely the making and assessment of verifiable claims about AI development through a variety of mechanisms. A richer toolbox of mechanisms for this purpose can inform developers' efforts to earn trust, the demands made of AI developers by activists and civil society organizations, and regulators' efforts to ensure that AI is developed responsibly.

Responsibility

Diligence

Ethics: If the widespread articulation of ethical principles can be seen as a first step toward ensuring responsible AI development, insofar as it helped to establish a standard against which behavior can be judged, then the adoption of mechanisms to make verifiable claims represents a second.

Verifiability: The authors of this report are eager to see further steps forward and hope that the framing of these mechanisms inspires the AI community to begin a meaningful dialogue around approaching verifiability in a collaborative fashion across organizations. We are keen to discover, study, and foreground additional institutional, software, and hardware mechanisms that could help enable trustworthy AI development. We encourage readers interested in collaborating in these or other areas to contact the corresponding authors of the report.

Continuous Improvement: As suggested by the title of the report (which references supporting verifiable claims rather than ensuring them), we see the mechanisms discussed here as enabling incremental improvements rather than providing a decisive solution to the challenge of verifying claims in the AI ecosystem. And despite the benefits associated with verifiable claims, they are also insufficient to ensure that AI developers will behave responsibly. There are at least three reasons for this.

Generality: First, there is a tension between verifiability of claims and the generality of such claims. This tension arises because the narrow properties of a system are easier to verify than the general ones, which tend to be of greater social interest. Safety writ large, for example, is inherently harder to verify than performance on a particular metric for safety. Additionally, broad claims about the beneficial societal impacts of a system or organization are harder to verify than more circumscribed claims about impacts in specific contexts.

Practicality: Second, the verifiability of claims does not ensure that they will be verified in practice. The mere existence of mechanisms for supporting verifiable claims does not ensure that they will be demanded by consumers, citizens, and policymakers (and even if they are, the burden ought not to be on them to do so). For example, consumers often use technologies in ways that are inconsistent with their stated values (e.g., a concern for personal privacy) because other factors such as convenience and brand loyalty also play a role in influencing their behavior [128].

Regulation: Third, even if a claim about AI development is shown to be false, asymmetries of power may prevent corrective steps from being taken. Members of marginalized communities, who often bear the brunt of harms associated with AI [2], often lack the political power to resist technologies that they deem detrimental to their interests. Regulation will be required to ensure that AI developers provide evidence that bears on important claims they make, to limit applications of AI where there is insufficient technical and social infrastructure for ensuring responsible

development, or to increase the variety of viable options available to consumers that are consistent with their stated values.

Progress: These limitations notwithstanding, verifiable claims represent a step toward a more trustworthy AI development ecosystem. Without a collaborative effort between AI developers and other stakeholders to improve the verifiability of claims, society's concerns about AI development are likely to grow: AI is being applied to an increasing range of high-stakes tasks, and with this wide deployment comes a growing range of risks. With a concerted effort to enable verifiable claims about AI development, there is a greater opportunity to positively shape AI's impact and increase the likelihood of widespread societal benefits.

Behavioral Standards: The creation and public announcement of a code of ethics proclaims an organization's commitment to ethical conduct both externally to the wider public, as well as internally to its employees, boards, and shareholders. Codes of conduct differ from codes of ethics in that they contain a set of concrete behavioral standards... Many organizations use the terms synonymously. The specificity of codes of ethics can vary, and more specific (i.e., action-guiding) codes of ethics (i.e. those equivalent to codes of conduct) can be better for earning trust because they are more falsifiable. Additionally, the form and content of these mechanisms can evolve over time–consider, e.g., Google's AI Principles, which have been incrementally supplemented with more concrete guidance in particular areas.

Transparency: Transparency measures could be undertaken on a voluntary basis or as part of an agreed framework involving relevant parties (such as a consortium of AI developers, interested non-profits, or policymakers). For example, algorithmic impact assessments are intended to support affected communities and stakeholders in assessing AI and other automated decision systems [2]. The Canadian government, for example, has centered AIAs in its Directive on Automated Decision-Making [25] [26]. Another path toward greater transparency around AI development involves increasing the extent and quality of documentation for AI systems. Such documentation can help foster informed and safe use of AI systems by providing information about AI systems' biases and other attributes [27][28][29].

Auditing: Auditing is a structured process by which an organization's present or past behavior is assessed for consistency with relevant principles, regulations, or norms. Auditing has promoted consistency and accountability in industries outside of AI such as finance and air travel. In each case, auditing is tailored to the evolving nature of the industry in question. Recently, auditing has gained traction as a potential paradigm for assessing whether AI development was conducted in a manner consistent with the stated principles of an organization, with valuable work focused on designing internal auditing processes (i.e. those in which the auditors are also employed by the organization being audited) [36].

Best Practices: Techniques and best practices have not yet been established for auditing AI systems. Outside of AI, however, there are well-developed frameworks on which to build. Outcomes- or claim-based "assurance frameworks" such as the Claims-Arguments-Evidence framework (CAE) and Goal Structuring Notation (GSN) are already in wide use in safety-critical auditing contexts.20 By allowing different types of arguments and evidence to be used appropriately by auditors, these frameworks provide considerable flexibility in how high-level claims are substantiated, a needed feature given the wide ranging and fast-evolving societal challenges posed by AI.

Privacy: Possible aspects of AI systems that could be independently audited include the level of privacy protection guaranteed, the extent to (and methods by) which the AI systems were tested for safety, security or ethical concerns, and the sources of data, labor, and other resources used.

Safety: Third party auditing could be applicable to a wide range of AI applications, as well. Safety-critical AI systems such as autonomous vehicles and medical AI systems, for example, could be audited for safety and security. Such audits could confirm or refute the accuracy of previous claims made by developers, or compare their efforts against an independent set of standards for safety and security.

Security

Fairness: As another example, search engines and recommendation systems could be independently audited for harmful biases.

Beneficence: Auditing imposes costs (financial and otherwise) that must be weighed against its value. Even if auditing is broadly societally beneficial and non-financial costs (e.g., to intellectual property) are managed, the financial costs

will need to be borne by someone (auditees, large actors in the industry, taxpayers, etc.), raising the question of how to initiate a self-sustaining process by which third party auditing could mature and scale. However, if done well, third party auditing could strengthen the ability of stakeholders in the AI ecosystem to make and assess verifiable claims. And notably, the insights gained from third party auditing could be shared widely, potentially benefiting stakeholders even in countries with different regulatory approaches for AI.

Collaboration: Doing red teaming in a more collaborative fashion, as a community of focused professionals across organizations, has several potential benefits: • Participants in such a community would gain useful, broad knowledge about the AI ecosystem, allowing them to identify common attack vectors and make periodic ecosystem-wide recommendations to organizations that are not directly participating in the core community; • Collaborative red teaming distributes the costs for such a team across AI developers, allowing those who otherwise may not have utilized a red team of similarly high quality or one at all to access its benefits (e.g., smaller organizations with less resources); • Greater collaboration could facilitate sharing of information about security-related AI incidents.

Incentives: Note that bounties are not sufficient for ensuring that a system is safe, secure, or fair, and it is important to avoid creating perverse incentives (e.g., encouraging work on poorly-specified bounties and thereby negatively affecting talent pipelines) [50]. Some system properties can be difficult to discover even with bounties, and the bounty hunting community might be too small to create strong assurances. However, relative to the status quo, bounties might increase the amount of scrutiny applied to AI systems.

Reproducibility: Reproducibility of technical results in AI is a key way of enabling verification of claims about system properties, and a number of ongoing initiatives are aimed at improving reproducibility in AI. Publication of results, models, and code increase the ability of outside parties (especially technical experts) to verify claims made about AI systems. Careful experimental design and the use of (and contribution to) standard software libraries can also improve reproducibility of particular results.

Formal Verification (Software): Formal verification establishes whether a system satisfies some requirements using the formal methods of mathematics. Formal verification is often a compulsory technique deployed in various safety-critical domains to provide guarantees regarding the functional behaviors of a system. These are typically guarantees that testing cannot provide. Until recently, AI systems utilizing machine learning (ML) have not generally been subjected to such rigor, but the increasing use of ML in safety-critical domains, such as automated transport and robotics, necessitates the creation of novel formal analysis techniques addressing ML models and their accompanying non-ML components. Techniques for formally verifying ML models are still in their infancy and face numerous challenges, which we discuss in Appendix VI(A).

Formal Verification (Hardware): Formal verification, discussed above in the software mechanisms section, is the process of establishing whether a software or hardware system satisfies some requirements or properties, using formal methods to generate mathematical proofs. Practical tools, such as GPUVerify for GPU kernels, exist to formally verify components of the AI hardware base, but verification of the complete hardware base is currently an ambitious goal. Because only parts of the AI hardware ecosystem are verified, it is important to map which properties are being verified for different AI accelerators and under what assumptions, who has access to evidence of such verification processes (which may be part of a third party audit), and what properties we should invest more research effort into verifying (or which assumption would be a priority to drop).

Empirical Verification: The empirical verification and validation of machine learning by machine learning has been proposed as an alternative paradigm to formal verification. Notably, it can be more practical than formal verification, but since it operates empirically, the method cannot as fully guarantee its claims. Machine learning could be used to search for common error patterns in another system's code, or be used to create simulation environments to adversarially find faults in an AI system's behavior. For example, adaptive stress testing (AST) of an AI system allows users to find the most likely failure of a system for a given scenario using reinforcement learning [61], and is being used by to validate the next generation of aircraft collision avoidance software [62]. Techniques requiring further research include using machine learning to evaluate another machine learning system (either by directly inspecting its policy or by creating environments to test the model) and using ML to evaluate the input of another machine learning model. In the future, data from model failures, especially pooled across multiple labs and stakeholders, could potentially be used to create classifiers that detect suspicious or anomalous AI behavior.

Practical Verification: Practical verification is the use of scientific protocols to characterize a model's data, assumptions, and performance. Training data can be rigorously evaluated for representativeness [63] [64]; assumptions can be characterized by evaluating modular components of an AI model and by clearly communicating output uncertainties; and performance can be characterized by measuring generalization, fairness, and performance heterogeneity across population subsets. Causes of differences in performance between models could be robustly attributed via randomized controlled trials.

Adversarial Robustness: A developer may wish to make claims about a system's adversarial robustness. 45 Currently, the security balance is tilted in favor of attacks rather than defenses, with only adversarial training [65] having stood the test of multiple years of attack research. Certificates of robustness, based on formal proofs, are typically approximate and give meaningful bounds of the increase in error for only a limited range of inputs, and often only around the data available for certification (i.e. not generalizing well to unseen data [66] [67] [68]). Without approximation, certificates are computationally prohibitive for all but the smallest real world tasks [69]. Further, research is needed on scaling formal certification methods to larger model sizes.

Attestation: Remote attestation leverages a "root of trust" (provided in hardware or in software, e.g., a secret key stored in isolated memory) to cryptographically sign a measurement or property of the system, thus providing a remote party proof of the authenticity of the measurement or property. Remote attestation is often used to attest that a certain version of software is currently running, or that a computation took a certain amount of time (which can then be compared to a reference by the remote party to detect tampering) [106].

1. Institutions

Ensure that individuals or organizations making claims regarding AI development are incentivized to be diligent in developing AI responsibly and that other stakeholders can verify that behavior.

Stakeholder(s)

Institutions :
Institutions may be formal and public institutions, such as: laws, courts, and regulatory agencies; private formal arrangements between parties, such as contracts; interorganizational structures such as industry associations, strategic alliances, partnerships, coalitions, joint ventures, and research consortia. Institutions may also be informal norms and practices that prescribe behaviors in particular contexts; or third party organizations, such as professional bodies and academic institutions.

Institutional Mechanisms and Recommendations — "Institutional mechanisms" are processes that shape or clarify the incentives of the people involved in AI development, make their behavior more transparent, or enable accountability for their behavior. Institutional mechanisms help to ensure that individuals or organizations making claims regarding AI development are incentivized to be diligent in developing AI responsibly and that other stakeholders can verify that behavior. Institutions can shape incentives or constrain behavior in various ways. Several clusters of existing institutional mechanisms are relevant to responsible AI development, and we characterize some of their roles and limitations below. These provide a foundation for the subsequent, more detailed discussion of several mechanisms and associated recommendations. Specifically, we provide an overview of some existing institutional mechanisms that have the following functions:

- Clarifying organizational goals and values;
- Increasing transparency regarding AI development processes;
- Creating incentives for developers to act in ways that are responsible; and
- Fostering exchange of information among developers.

1.1. Goals & Values

Clarify the goals and values.

Institutional mechanisms can help clarify an organization's goals and values, which in turn can provide a basis for evaluating their claims. These statements of goals and values–which can also be viewed as (high level) claims in the framework discussed here–can help to contextualize the actions an organization takes and lay the foundation for others (shareholders, employees, civil society organizations, governments, etc.) to monitor and evaluate behavior. Over 80 AI organizations [5], including technology companies such as Google [22], OpenAI [23], and Microsoft [24] have publicly stated the principles they will follow in developing AI. Codes of ethics or conduct are far from sufficient, since they are typically abstracted away from particular cases and are not reliably enforced, but they can be valuable by establishing criteria that a developer concedes are appropriate for evaluating its behavior. The creation and public announcement of a code of ethics proclaims an organization's commitment to ethical conduct both externally to the wider public, as well as internally to its employees, boards, and shareholders. Codes of conduct differ from codes of ethics in that they contain a set of concrete behavioral standards.

1.2. Transparency

Enable others to more easily verify compliance with appropriate norms, regulations, and agreements.

Institutional mechanisms can increase transparency regarding an organization's AI development processes in order to permit others to more easily verify compliance with appropriate norms, regulations, or agreements. Improved transparency may reveal the extent to which actions taken by an AI developer are consistent with their declared intentions and goals. The more reliable, timely, and complete the institutional measures to enhance transparency are, the more assurance may be provided.

1.3. Incentives

Create incentives for organizations to act in ways that are responsible.

Institutional mechanisms can create incentives for organizations to act in ways that are responsible. Incentives can be created within an organization or externally, and they can operate at an organizational or an individual level. The incentives facing an actor can provide evidence regarding how that actor will behave in the future, potentially bolstering the credibility of related claims. To modify incentives at an organizational level, organizations can choose to adopt different organizational structures (such as benefit corporations) or take on legally binding intra-organizational commitments. For example, organizations could credibly commit to distributing the benefits of AI broadly through a legal commitment that shifts fiduciary duties. Institutional commitments to such steps could make a particular organization's financial incentives more clearly aligned with the public interest. To the extent that commitments to responsible AI development and distribution of benefits are widely implemented, AI developers would stand to benefit from each others' success, potentially reducing incentives to race against one another [1]. And critically, government regulations such as the General Data Protection Regulation (GDPR) enacted by the European Union shift developer incentives by imposing penalties on developers that do not adequately protect privacy or provide recourse for algorithmic decision-making.

1.4. Information

Foster exchange of information between developers.

Finally, institutional mechanisms can foster exchange of information between developers. To avoid "races to the bottom" in AI development, AI developers can exchange lessons learned and demonstrate their compliance with relevant norms to one another. Multilateral fora (in addition to bilateral conversations between organizations) provide opportunities for discussion and repeated interaction, increasing transparency and interpersonal understanding. Voluntary membership organizations with stricter rules and norms have been implemented in other industries and might also be a useful model for AI developers [31]. Steps in the direction of robust information exchange between AI developers include the creation of consensus around important priorities such as safety, security, privacy, and fairness; participation in multi-stakeholder fora such as the Partnership on Artificial Intelligence to Benefit People and Society (PAI), the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE), the International Telecommunications Union (ITU), and the International Standards Organization (ISO); and clear identification of roles or offices within organizations who are responsible for maintaining and deepening interorganizational communication [10].

Stakeholder(s):

AI Developers

1.4.1. Priorities

Create consensus around important priorities such as safety, security, privacy, and fairness.

1.4.2. Multi-Stakeholder Fora

Participate in multi-stakeholder fora.

Stakeholder(s):

Partnership on Artificial Intelligence to Benefit People and Society (PAI)

Association for Computing Machinery (ACM)

Institute of Electrical and Electronics Engineers (IEEE)

International Telecommunications Union (ITU)

International Standards Organization (ISO)

1.4.3. Communication

Clearly identify roles or offices within organizations who are responsible for maintaining and deepening interorganizational communication.

1.4.4. Incentives

Examine the incentives (and disincentives) for free flow of information within an organization.

It is also important to examine the incentives (and disincentives) for free flow of information within an organization. Employees within organizations developing AI systems can play an important role in identifying unethical or unsafe practices. For this to succeed, employees must be well-informed about the scope of AI development efforts within their organization and be comfortable raising their concerns, and such concerns need to be taken seriously by management. Policies (whether governmental or organizational) that help ensure safe channels for expressing concerns are thus key foundations for verifying claims about AI development being conducted responsibly.

Recommendation 1. Auditing

Research options for conducting and funding third party auditing of AI systems.

A coalition of stakeholders should create a task force to research options for conducting and funding third party auditing of AI systems. — Problem: The process of AI development is often opaque to those outside a given organization, and various barriers make it challenging for third parties to verify the claims being made by a developer. As a result, claims about system attributes may not be easily verified... Auditing could take at least four quite different forms, and likely further variations are possible: auditing by an independent body with government-backed policing and sanctioning power; auditing that occurs entirely within the context of a government, though with multiple agencies involved [37]; auditing by a private expert organization or some ensemble of such organizations; and internal auditing followed by public disclosure of (some subset of) the results. As commonly occurs in other contexts, the results produced by independent auditors might be made publicly available, to increase confidence in the propriety of the auditing process.

Stakeholder(s):

AI Developers :
AI developers have justifiable concerns about being transparent with information concerning commercial secrets, personal information, or AI systems that could be misused; however, problems arise when these concerns incentivize them to evade scrutiny. Third party auditors can be given privileged and secured access to this private information, and they can be tasked with assessing whether safety, security, privacy, and fairness-related claims made by the AI developer are accurate.

Third-Party Auditors :
Third party auditing is a form of auditing conducted by an external and independent auditor, rather than the organization being audited, and can help address concerns about the incentives for accuracy in self-reporting. Provided that they have sufficient information about the activities of an AI system, independent auditors with strong reputational and professional incentives for truthfulness can help verify claims about AI development.

Governments :
Third party auditors should be held accountable by government, civil society, and other stakeholders to ensure that strong incentives exist to act accurately and fairly. Reputational considerations help to ensure auditing integrity in the case of financial accounting, where firms prefer to engage with credible auditors [38].

Civil Society

Licensing Agencies :
Alternatively, a licensing system could be implemented in which auditors undergo a standard training process in order to become a licensed AI system auditor. However, given the variety of methods and applications in the field of AI, it is not obvious whether auditor licensing is a feasible option for the industry: perhaps a narrower form of licensing would be helpful (e.g., a subset of AI such as adversarial machine learning).

AI Auditing Task Force :
AI developers and other stakeholders (such as civil society organizations and policymakers) should collaboratively explore the challenges associated with third party auditing. A task force focused on this issue could explore appropriate initial domains/applications to audit, devise approaches for handling sensitive intellectual property, and balance the need for standardization with the need for flexibility as AI technology evolves. Collaborative research into this domain seems especially promising given that the same auditing process could be used across labs and countries. As research in these areas evolves, so too will auditing processes–one might thus think of auditing as a "meta-mechanism" which could involve assessing the quality of other efforts discussed in this report such as red teaming.

Regulators :
One way that third party auditing could connect to government policies, and be funded, is via a "regulatory market" [42]. In a regulatory market for AI, a government would establish high-level outcomes to be achieved from regulation of AI (e.g., achievement of a certain level of safety in an industry) and then create or support private sector entities or other organizations that compete in order to design and implement the precise technical oversight required to achieve those outcomes. Regardless of whether such an approach is pursued, third party auditing by private actors should be viewed as a complement to, rather than a substitute, for governmental regulation. And regardless of the entity conducting oversight of AI developers, in any case there will be a need to grapple with difficult challenges such as the treatment of proprietary data.

Recommendation 2. Red Teams

Run red teaming exercises to explore risks associated with AI systems.

Organizations developing AI should run red teaming exercises to explore risks associated with systems they develop, and should share best practices and tools for doing so. — Problem: It is difficult for AI developers to address the "unknown unknowns" associated with AI systems, including limitations and risks that might be exploited by malicious actors. Further, existing red teaming approaches are insufficient for addressing these concerns in the AI context.

Stakeholder(s):

AI Developers :
In order for AI developers to make verifiable claims about their AI systems being safe or secure, they need processes for surfacing and addressing potential safety and security risks. Practices such as red teaming exercises help organizations to discover their own limitations and vulnerabilities as well as those of the AI systems they develop, and to approach them holistically, in a way that takes into account the larger environment in which they are operating.

Red Teams :
A red team exercise is a structured effort to find flaws and vulnerabilities in a plan, organization, or technical system, often performed by dedicated "red teams" that seek to adopt an attacker's mindset and methods. In domains such as computer security, red teams are routinely tasked with emulating attackers in order to find flaws and vulnerabilities in organizations and their systems. Discoveries made by red teams allow organizations to improve security and system integrity before and during deployment. Knowledge that a lab has a red team can potentially improve the trustworthiness of an organization with respect to their safety and security claims, at least to the extent that effective red teaming practices exist and are demonstrably employed.

Attack Teams :
As indicated by the number of cases in which AI systems cause or threaten to cause harm, developers of an AI system often fail to anticipate the potential risks associated with technical systems they develop. These risks include both inadvertent failures and deliberate misuse. Those not involved in the development of a particular system may be able to more easily adopt and practice an attacker's skillset. A growing number of industry labs have dedicated red teams, although best practices for such efforts are generally in their early stages. There is a need for experimentation both within and across organizations in order to move red teaming in AI forward, especially since few AI developers have expertise in relevant areas such as threat modeling and adversarial machine learning [44].

AI Red Teaming Professionals :
AI systems and infrastructure vary substantially in terms of their properties and risks, making in-house red-teaming expertise valuable for organizations with sufficient resources. However, it would also be beneficial to experiment with the formation of a community of AI red teaming professionals that draws together individuals from different organizations and backgrounds, specifically focused on some subset of AI (versus AI in general) that is relatively well-defined and relevant across multiple organizations. A community of red teaming professionals could take actions such as publish best practices, collectively analyze particular case studies, organize workshops on emerging issues, or advocate for policies that would enable red teaming to be more effective.

Red Teaming Community :
Two critical questions that would need to be answered in the context of forming a more cohesive AI red teaming community are: what is the appropriate scope of such a group, and how will proprietary information be handled? The two questions are related. Particularly competitive contexts (e.g., autonomous vehicles) might be simultaneously very appealing and challenging: multiple parties stand to gain from pooling of insights, but collaborative red teaming in such contexts is also challenging because of intellectual property and security concerns. As an alternative to or supplement to explicitly collaborative red teaming, organizations building AI technologies should establish shared resources and outlets for sharing relevant non-proprietary information. The subsection on sharing of AI incidents also discusses some potential innovations that could alleviate concerns around sharing proprietary information.

Recommendation 3. Bounties

Pilot bias and safety bounties for AI systems.

Problem: There is too little incentive, and no formal process, for individuals unaffiliated with a particular AI developer to seek out and report problems of AI bias and safety. As a result, broad-based scrutiny of AI systems for these properties is relatively rare... Issues to be addressed in setting up such a bounty program include [46]: • Setting compensation rates for different scales/severities of issues discovered; • Determining processes for soliciting and evaluating bounty submissions; • Developing processes for disclosing issues discovered via such bounties in a timely fashion; • Designing appropriate interfaces for reporting of bias and safety problems in the context of deployed AI systems; • Defining processes for handling reported bugs and deploying fixes; • Avoiding creation of perverse incentives.

There is not a perfect analogy between discovering and addressing traditional computer security vulnerabilities, on the one hand, and identifying and addressing limitations in AI systems, on the other. Work is thus needed to explore the factors listed above in order to adapt the bug bounty concept to the context of AI development. The computer security community has developed norms (though not a consensus) regarding how to address "zero day" vulnerabilities, but no comparable norms yet exist in the AI community. There may be a need for distinct approaches to different types of vulnerabilities and associated bounties, depending on factors such as the potential for remediation of the issue and the stakes associated with the AI system. Bias might be treated differently from safety issues such as unsafe exploration, as these have distinct causes, risks, and remediation steps. In some contexts, a bounty might be paid for information even if there is no ready fix to the identified issue, because providing accurate documentation to system users is valuable in and of itself and there is often no pretense of AI systems being fully robust. In other cases, more care will be needed in responding to the identified issue, such as when a model is widely used in deployed products and services.

Stakeholder(s):

AI Developers :
AI developers should pilot bias and safety bounties for AI systems to strengthen incentives and processes for broad-based scrutiny of AI systems.

Information Security Industry :
"Bug bounty" programs have been popularized in the information security industry as a way to compensate individuals for recognizing and reporting bugs, especially those related to exploits and vulnerabilities [45]. Bug bounties provide a legal and compelling way to report bugs directly to the institutions affected, rather than exposing the bugs publicly or selling the bugs to others. Typically, bug bounties involve an articulation of the scale and severity of the bugs in order to determine appropriate compensation.

AI Bug Bounty Hunters :
While efforts such as red teaming are focused on bringing internal resources to bear on identifying risks associated with AI systems, bounty programs give outside individuals a method for raising concerns about specific AI systems in a formalized way. Bounties provide one way to increase the amount of scrutiny applied to AI systems, increasing the likelihood of claims about those systems being verified or refuted. Bias28 and safety bounties would extend the bug bounty concept to AI, and could complement existing efforts to better document datasets and models for their performance limitations and other properties. We focus here on bounties for discovering bias and safety issues in AI systems as a starting point for analysis and experimentation, but note that bounties for other properties (such as security, privacy protection, or interpretability) could also be explored.

Ziad Obermeyer :
While some instances of bias are easier to identify, others can only be uncovered with significant analysis and resources. For example, Ziad Obermeyer et al. uncovered racial bias in a widely used algorithm affecting millions of patients [47].

Bias Hunters

Consumers :
There have also been several instances of consumers with no direct access to AI institutions using social media and the press to draw attention to problems with AI [48].

Social Media

Investigative Journalists :
To date, investigative journalists and civil society organizations have played key roles in surfacing different biases in deployed AI systems.

Civil Society Organizations

Companies :
If companies were more open earlier in the development process about possible faults, and if users were able to raise (and be compensated for raising) concerns about AI to institutions, users might report them directly instead of seeking recourse in the court of public opinion.

— continued next page

Stakeholders (continued)

Safety Risk Hunters :
In addition to bias, bounties could also add value in the context of claims about AI safety. Algorithms or models that are purported to have favorable safety properties, such as enabling safe exploration or robustness to distributional shifts [49], could be scrutinized via bounty programs. To date, more attention has been paid to documentation of models for bias properties than safety properties, though in both cases, benchmarks remain in an early state. Improved safety metrics could increase the comparability of bounty programs and the overall robustness of the bounty ecosystem; however, there should also be means of reporting issues that are not well captured by existing metrics.

Recommendation 4. Incidents

Share information about AI incidents.

Problem: Claims about AI systems can be scrutinized more effectively if there is common knowledge of the potential risks of such systems. However, cases of desired or unexpected behavior by AI systems are infrequently shared since it is costly to do unilaterally... The sharing of AI incidents can improve the verifiability of claims in AI development by highlighting risks that might not have otherwise been considered by certain actors. Knowledge of these risks, in turn, can then be used to inform questions posed to AI developers, increasing the effectiveness of external scrutiny. Incident sharing can also (over time, if used regularly) provide evidence that incidents are found and acknowledged by particular organizations, though additional mechanisms would be needed to demonstrate the completeness of such sharing... Improving the ability and incentive of AI developers to report incidents requires building additional infrastructure, analogous to the infrastructure that exists for reporting incidents in other domains such as cybersecurity. Infrastructure to support incident sharing that involves non-public information would require the following resources: • Transparent and robust processes to protect organizations from undue reputational harm brought about by the publication of previously unshared incidents. This could be achieved by anonymizing incident information to protect the identity of the organization sharing it. Other information-sharing methods should be explored that would mitigate reputational risk to organizations, while preserving the usefulness of information shared; • A trusted neutral third party that works with each organization under a non-disclosure agreement to collect and anonymize private information; • An organization that maintains and administers an online platform where users can easily access the incident database, including strong encryption and password protection for private incidents as well as a way to submit new information. This organization would not have to be the same as the third party that collects and anonymizes private incident data; • Resources and channels to publicize the existence of this database as a centralized resource, to accelerate both contributions to the database and positive uses of the knowledge from the database; and • Dedicated researchers who monitor incidents in the database in order to identify patterns and shareable lessons.

The costs of incident sharing (e.g., public relations risks) are concentrated on the sharing organization, although the benefits are shared broadly by those who gain valuable information about AI incidents. Thus, a cooperative approach needs to be taken for incident sharing that addresses the potential downsides. A more robust infrastructure for incident sharing (as outlined above), including options for anonymized reporting, would help ensure that fear of negative repercussions from sharing does not prevent the benefits of such sharing from being realized.

Stakeholder(s):

AI Developers :
AI developers should share more information about AI incidents, including through collaborative channels... Developers should seek to share AI incidents with a broad audience so as to maximize their usefulness, and take advantage of collaborative channels such as centralized incident databases as that infrastructure matures. In addition, they should move towards publicizing their commitment to (and procedures for) doing such sharing in a routine way rather than in an ad-hoc fashion, in order to strengthen these practices as norms within the AI development community

Organizations :
Organizations can share AI "incidents," or cases of undesired or unexpected behavior by an AI system that causes or could cause harm, by publishing case studies about these incidents from which others can learn. This can be accompanied by information about how they have worked to prevent future incidents based on their own and others' experiences. By default, organizations developing AI have an incentive to primarily or exclusively report positive outcomes associated with their work rather than incidents. As a result, a skewed image is given to the public, regulators, and users about the potential risks associated with AI development.

The Public

Regulators

AI Users

Partnership on AI :
AI incidents can include those that are publicly known and transparent, publicly known and anonymized, privately known and anonymized, or privately known and transparent. The Partnership on AI has begun building an AI incident-sharing database, called the AI Incident Database. The pilot was built using publicly available information through a set of volunteers and contractors manually collecting

— continued next page

Stakeholders (continued)

known AI incidents where AI caused harm in the real world.

Cybersecurity Community :
Incident sharing is closely related to but distinct from responsible publication practices in AI and coordinated disclosure of cybersecurity vulnerabilities [55]. Beyond implementation of progressively more robust platforms for incident sharing and contributions to such platforms, future work could also explore connections between AI and other domains in more detail, and identify key lessons from other domains in which incident sharing is more mature (such as the nuclear and cybersecurity industries).

Governments :
Over the longer term, lessons learned from experimentation and research could crystallize into a mature body of knowledge on different types of AI incidents, reporting processes, and the costs associated with incident sharing. This, in turn, can inform any eventual government efforts to require or incentivize certain forms of incident reporting.

2. Software

Software Mechanisms and Recommendations — Software mechanisms involve shaping and revealing the functionality of existing AI systems. They can support verification of new types of claims or verify existing claims with higher confidence. This section begins with an overview of the landscape of software mechanisms relevant to verifying claims, and then highlights several key problems, mechanisms, and associated recommendations. Software mechanisms, like software itself, must be understood in context (with an appreciation for the role of the people involved). Expertise about many software mechanisms is not widespread, which can create challenges for building trust through such mechanisms. For example, an AI developer that wants to provide evidence for the claim that "user data is kept private" can help build trust in the lab's compliance with a a formal framework such as differential privacy, but non-experts may have in mind a different definition of privacy. It is thus critical to consider not only which claims can and can't be substantiated with existing mechanisms in theory, but also who is well-positioned to scrutinize these mechanisms in practice. Keeping their limitations in mind, software mechanisms can substantiate claims associated with AI development in various ways that are complementary to institutional and hardware mechanisms. They can allow researchers, auditors, and others to understand the internal workings of any given system. They can also help characterize the behavioral profile of a system over a domain of expected usage. Software mechanisms could support claims such as:

- This system is robust to 'natural' distributional shifts [49] [56];
- This system is robust even to adversarial examples [57] [58];
- This system has a well-characterized error surface and users have been informed of contexts in which the system would be unsafe to use;
- This system's decisions exhibit statistical parity with respect to sensitive demographic attributes; and
- This system provides repeatable or reproducible results.

Recommendation 5. Audit Trails

Develop audit trail requirements for safety-critical applications.

Problem: AI systems lack traceable logs of steps taken in problem-definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those systems' properties and impacts. Audit trails can improve the verifiability of claims about engineered systems, although they are not yet a mature mechanism in the context of AI. An audit trail is a traceable log of steps in system operation, and potentially also in design and testing. We expect that audit trails will grow in importance as AI is applied to more safety-critical contexts. They will be crucial in supporting many institutional trust-building mechanisms, such as third-party auditors, government regulatory bodies, and voluntary disclosure of safety-relevant information by companies. Audit trails could cover all steps of the AI development process, from the institutional work of problem and purpose definition leading up to the initial creation of a system, to the training and development of that system, all the way to retrospective accident analysis.

Stakeholder(s):

Standards Setting Bodies :
Standards setting bodies should work with academia and industry to develop audit trail requirements for safety-critical applications of AI systems. — Organizations involved in setting technical standards–including governments and private actors–should establish clear guidance regarding how to make safety-critical AI systems fully auditable. Although application dependent, software audit trails often require a base set of traceability trails to be demonstrated for qualification; the decision to choose a certain set of trails requires considering trade-offs about efficiency, completeness, tamper-proofing, and other design considerations. There is flexibility in the type of documents or evidence the auditee presents to satisfy these general traceability requirements (e.g., between test logs and requirement documents, verification and validation activities, etc.). Existing standards often define in detail the required audit trails for specific applications.

IEC :
For example, IEC 61508 is a basic functional safety standard required by many industries, including nuclear power. Such standards are not yet established for AI systems. A wide array of audit trails related to an AI development process can already be produced, such as code changes, logs of training runs, all outputs of a model, etc. Inspiration might be taken

— continued next page

Stakeholders (continued)

from recent work on internal algorithmic auditing [36] and ongoing work on the documentation of AI systems more generally, such as the ABOUT ML project [27]. Importantly, we recommend that in order to have maximal impact, any standards for AI audit trails should be published freely, rather than requiring payment as is often the case.

Academia

Industry :
There is already strong precedence for audit trails in numerous industries, in particular for safety-critical systems. Commercial aircraft, for example, are equipped with flight data recorders that record and capture multiple types of data each second [70]. In safety-critical domains, the compliance of such evidence is usually assessed within a larger "assurance case" utilising the CAE or Goal-Structuring-Notation (GSN) frameworks. Tools such as the Assurance and Safety Case Environment (ACSE) exist to help both the auditor and the auditee manage compliance claims and corresponding evidence. Version control tools such as GitHub or GitLab can be utilized to demonstrate individual document traceability. Proposed projects like Verifiable Data Audit [71] could establish confidence in logs of data interactions and usage.

Recommendation 6. Interpretability

Support research into the interpretability of AI systems.

Organizations developing AI and funding bodies should support research into the interpretability of AI systems, with a focus on supporting risk assessment and auditing. — Problem: It's difficult to verify claims about "black-box" AI systems that make predictions without explanations or visibility into their inner workings. This problem is compounded by a lack of consensus on what interpretability means. Despite remarkable performance on a variety of problems, AI systems are frequently termed "black boxes" due to the perceived difficulty of understanding and anticipating their behavior. This lack of interpretability in AI systems has raised concerns about using AI models in high stakes decision-making contexts where human welfare may be compromised [73]. Having a better understanding of how the internal processes within these systems work can help proactively anticipate points of failure, audit model behavior, and inspire approaches for new systems. Research in model interpretability is aimed at helping to understand how and why a particular model works. A precise, technical definition for interpretability is elusive; by nature, the definition is subject to the inquirer. Characterizing desiderata for interpretable models is a helpful way to formalize interpretability [74] [75]. Useful interpretability tools for building trust are also highly dependent on the target user and the downstream task. For example, a model developer or regulator may be more interested in understanding model behavior over the entire input distribution whereas a novice layperson may wish to understand why the model made a particular prediction for their individual case. Crucially, an "interpretable" model may not be necessary for all situations. The weight we place upon a model being interpretable may depend upon a few different factors, for example: • More emphasis in sensitive domains (e.g., autonomous driving or healthcare, where an incorrect prediction adversely impacts human welfare) or when it is important for end-users to have actionable recourse (e.g., bank loans) [77]; • Less emphasis given historical performance data (e.g., a model with sufficient historical performance may be used even if it's not interpretable); and • Less emphasis if improving interpretability incurs other costs (e.g., compromising privacy).

In the longer term, for sensitive domains where human rights and/or welfare can be harmed, we anticipate that interpretability will be a key component of AI system audits, and that certain applications of AI will be gated on the success of providing adequate intuition to auditors about the model behavior. This is already the case in regulated domains such as finance [78]. An ascendent topic of research is how to compare the relative merits of different interpretability methods in a sensible way. Two criteria appear to be crucial: a. The method should provide sufficient insight for the end-user to understand how the model is making its predictions (e.g., to assess if it aligns with human judgment), and b. the interpretable explanation should be faithful to the model, i.e., accurately reflect its underlying behavior. Work on evaluating a., while limited in treatment, has primarily centered on comparing methods using human surveys [80]. More work at the intersection of human-computer interaction, cognitive science, and interpretability research–e.g., studying the efficacy of interpretability tools or exploring possible interfaces–would be welcome, as would further exploration of how practitioners currently use such tools [81] [82] [83] [78] [84]. Evaluating b., the reliability of existing methods is an active area of research [85] [86] [87] [88] [89] [90] [91] [92] [93]. This effort is complicated by the lack of ground truth on system behavior (if we could reliably anticipate model behavior under all circumstances, we would not need an interpretability method). The wide use of interpretable tools in sensitive domains underscores the continued need to develop benchmarks that assess the reliability of produced model explanations. It is important that techniques developed under the umbrella of interpretability not be used to provide clear explanations when such clarity is not feasible. Without sufficient rigor, interpretability could be used in service of unjustified trust by providing misleading explanations for system behavior. In identifying, carrying out, and/or funding research on interpretability, particular attention should be paid to whether and how such research might eventually aid in verifying claims about AI systems with high degrees of confidence to support risk assessment and auditing... Some areas of interpretability research are more developed than others. For example, attribution methods for explaining individual predictions of computer vision models are arguably one of the most welldeveloped research areas. As such, we suggest that the following under-explored directions would be useful for the development of interpretability tools that could support verifiable claims about system properties:

6.1. Criteria, Objectives & Frameworks

Establish consensus on the criteria, objectives, and frameworks for interpretability research.

Developing and establishing consensus on the criteria, objectives, and frameworks for interpretability research;

6.2. Provenance

Study the provenance of learned models.

Studying the provenance of a learned model (e.g., as a function of the distribution of training data, choice of particular model families, or optimization) instead of treating models as fixed; and

6.3. Constraints

Constrain models to be interpretable by default.

Constraining models to be interpretable by default, in contrast to the standard setting of trying to interpret a model post-hoc. This list is not intended to be exhaustive, and we recognize that there is uncertainty about which research directions will ultimately bear fruit. We discuss the landscape of interpretability research further in Appendix VI(C).

Recommendation 7. Privacy

Develop, share, and use suites of tools for privacy-preserving machine learning.

Problem: A range of methods can potentially be used to verifiably safeguard the data and models involved in AI development. However, standards are lacking for evaluating new privacy-preserving machine learning techniques, and the ability to implement them currently lies outside a typical AI developer's skill set. Training datasets for AI often include sensitive information about people, raising risks of privacy violation. These risks include unacceptable access to raw data (e.g., in the case of an untrusted employee or a data breach), unacceptable inference from a trained model (e.g., when sensitive private information can be extracted from a model), or unacceptable access to a model itself (e.g., when the model represents personalized preferences of an individual or is protected by intellectual property). For individuals to trust claims about an ML system sufficiently so as to participate in its training, they need evidence about data access (who will have access to what kinds of data under what circumstances), data usage, and data protection. The AI development community, and other relevant communities, have developed a range of methods and mechanisms to address these concerns, under the general heading of "privacy-preserving machine learning" (PPML) [94]. Privacy-preserving machine learning aims to protect the privacy of data or models used in machine learning, at training or evaluation time and during deployment. PPML has benefits for model users, and for those who produce the data that models are trained on. PPML is heavily inspired by research from the cryptography and privacy communities and is performed in practice using a combination of techniques, each with its own limitations and costs. These techniques are a powerful tool for supporting trust between data owners and model users, by ensuring privacy of key information. However, they must be used judiciously, with informed trade-offs among (1) privacy benefits, (2) model quality, (3) AI developer experience and productivity, and (4) overhead costs such as computation, communication, or energy consumption. They are also not useful in all contexts; therefore, a combination of techniques may be required in some contexts to protect data and models from the actions of well-resourced malicious actors. Before turning to our recommendation, we provide brief summaries of several PPML techniques that could support verifiable claims. Federated learning is a machine learning technique where many clients (e.g., mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g., service provider), while keeping the training data decentralized [95]. Each client's raw data is stored locally and not exchanged or transferred [95]. Federated learning addresses privacy concerns around the centralized collection of raw data, by keeping the data where it is generated (e.g., on the user's device or in a local silo) and only allowing model updates to leave the client. Federated learning does not, however, fully guarantee the privacy of sensitive data on its own, as some aspects of raw data could be memorized in the training process and extracted from the trained model if measures are not taken to address this threat. These measures include quantifying the degree to which models memorize training data [96], and incorporating differential privacy techniques to limit the contribution of individual clients in the federated setting [97]. Even when used by itself, federated learning addresses the threats that are endemic to centralized data collection and access, such as unauthorized access, data hacks, and leaks, and the inability of data owners to control their data lifecycle. Differential privacy [98] is a system for publicly sharing information derived from a dataset by describing the patterns of groups within the dataset, while withholding information about individuals in the dataset; it allows for precise measurements of privacy risks for current and potential data owners, and can address the raw-data-extraction threat described above. Differential privacy works through the addition of a controlled amount of statistical noise to obscure the data contributions from records or individuals in the dataset. Differential privacy is already used in various private and public AI settings, and researchers are exploring its role in compliance with new privacy regulations [100] [99]. Differential privacy and federated learning complement each other in protecting the privacy of raw data: federated learning keeps the raw data on the personal device, so it is never seen by the model trainer, while differential privacy ensures the model sufficiently prevents the memorization of raw data, so that it cannot be extracted from the model by its users. These techniques do not, however, protect the model itself from theft [101]. Encrypted computation addresses this risk by allowing the model to train and run on encrypted data while in an encrypted state, at the cost of overhead in terms of computation and communication. As a result, those training the model will not be able to see, leak, or otherwise abuse the data in its unencrypted form. The most well known methods for encrypted computation are homomorphic encryption, secure multi-party computation, and functional encryption [102]. For example, one of OpenMined's upcoming projects is Encrypted Machine Learning as a Service, which allows a model owner and

data owner to use their model and data to make a prediction, without the model owner disclosing their model, and without the data owner disclosing their data. These software mechanisms can guarantee tighter bounds on AI model usage than the legal agreements that developers currently employ, and tighter bounds on user data usage than institutional mechanisms such as user privacy agreements. Encrypted computation could also potentially improve the verifiability of claims by allowing sensitive models to be shared for auditing in a more secure fashion. A hardware-based method to protect models from theft (and help protect raw data from adversaries) is the use of secure enclaves, as discussed in Section 4.1 below. In the future, it may be possible to rely on a platform that enables verifiable data policies which address some of the security and privacy vulnerabilities in existing IT systems. One proposal for such a platform is Google's Project Oak, which leverages open source secure enclaves (see Section 4.1) and formal verification to technically enforce and assure policies around data storage, manipulation, and exchange. As suggested by this brief overview of PPML techniques, there are many opportunities for improving the privacy and security protections associated with ML systems. However, greater standardization of of PPML techniques–and in particular, the use of open source PPML frameworks that are benchmarked against common performance measures–may be needed in order for this to translate into a major impact on the verifiability of claims about AI development. First, robust open source frameworks are needed in order to reduce the skill requirement for implementing PPML techniques, which to date have primarily been adopted by large technology companies with in-house expertise in both ML and cryptography. Second, common standards for evaluating new PPML techniques could increase the comparability of new results, potentially accelerating research progress. Finally, standardization could improve the ability of external parties (including users, auditors, and policymakers) to verify claims about PPML performance.

7.1. PPML

Contribute to, use, and support the work of open-source communities working on PPML.

Where possible, AI developers should contribute to, use, and otherwise support the work of open-source communities working on PPML, such as OpenMined, Microsoft SEAL, tf-encrypted, tf-federated, and nGraph-HE. These communities have opened up the ability to use security and privacy tools in the ML setting, and further maturation of the software libraries built by these communities could yield still further benefits.

Stakeholder(s):

Open-Source Communities

7.2. Funding & Independence

Provide stable funding and ensure independence.

Open-source communities projects or projects backed by a particular company can sometimes suffer from a lack of stable funding support or independence as organizational priorities shift, suggesting a need for an AI community-wide approach to supporting PPML's growth. Notwithstanding some challenges associated with open source projects, they are uniquely amenable to broad-based scrutiny and iteration, and have yielded benefits already. Notably, integrated libraries for multiple techniques in privacy-preserving ML have started being developed for major deep learning frameworks such as TensorFlow and PyTorch.

7.3. Benchmarks

Develop benchmarks for PPML to unify goals and measure progress.

Benchmarks for PPML could help unify goals and measure progress across different groups.

7.4. Implementations

Maintain a repository of real-world implementation cases.

A centralized repository of real-world implementation cases, a compilation of implementation guides, and work on standardization/interoperability would all also aid in supporting adoption and scrutiny of privacy-preserving methods.

7.5. Guidance

Compile implementation guides.

7.6. Standardization & Interoperability

Work on standardization/interoperability.

3. Hardware

Stakeholder(s)

Cloud Service Providers :

Cloud computing: Hardware is also at the heart of the relationship between cloud providers and cloud users (as hardware resources are being rented). Associated verification mechanisms can help ensure that computations are being performed as promised, without the client having direct physical access to the hardware. For example, one could have assurances that a cloud-based AI service is not skimping on computations by running a less powerful model than advertised, operating on private data in a disallowed fashion, or compromised by malware [107]. Cloud providers are a promising intervention point for trust-building mechanisms; a single cloud provider services, and therefore has influence over, many customers.

AI Labs :

Even large AI labs rely predominantly on cloud computing for some or all of their AI development. Cloud providers already employ a variety of mechanisms to minimize risks of misuse on their platforms, including "Know Your Customer" services and Acceptable Use Policies. These mechanisms could be extended to cover AI misuse [108]. Additional mechanisms could be developed such as a forum where cloud providers can share best-practices about detecting and responding to misuse and abuse of AI through their services.

Hardware Mechanisms and Recommendations — Computing hardware enables the training, testing, and use of AI systems. Hardware relevant to AI development ranges from sensors, networking, and memory, to, perhaps most crucially, processing power [103]. Concerns about the security and other properties of computing hardware, as well as methods to address those concerns in a verifiable manner, long precede the current growth in adoption of AI. However, because of the increasing capabilities and impacts of AI systems and the particular hardware demands of the field, there is a need for novel approaches to assuring the verifiability of claims about the hardware used in AI development. Hardware mechanisms involve physical computing resources (e.g., CPUs and GPUs), including their distribution across actors, the ways they are accessed and monitored, and their properties (e.g., how they are designed, manufactured, or tested). Hardware can support verifiable claims in various ways. Secure hardware can play a key role in private and secure machine learning by translating privacy constraints and security guarantees into scrutable hardware designs or by leveraging hardware components in a software mechanism. Hardware mechanisms can also be used to demonstrate the ways in which an organization is using its general-purpose computing capabilities. At a higher level, the distribution of computing power across actors can potentially influence who is in a position to verify certain claims about AI development. This is true on the assumption that, all things being equal, more computing power will enable more powerful AI systems to be built, and that a technical capability to verify claims may itself require non-negligible computing resources. The use of standardized, publicly available hardware (sometimes called "commodity hardware") across AI systems also aids in the independent reproducibility of technical results, which in turn could play a role in technical auditing and other forms of accountability. Finally, hardware mechanisms can be deployed to enforce and verify policies relating to the security of the hardware itself (which, like software, might be compromised through error or malice).

Recommendation 8. Security

Develop hardware security features for AI accelerators and establish best practices for the use of hardware in machine learning contexts.

Industry and academia should work together to develop hardware security features for AI accelerators or otherwise establish best practices for the use of secure hardware (including secure enclaves on commodity

hardware) in machine learning contexts. — Problem: Hardware security features can provide strong assurances against theft of data and models, but secure enclaves (also known as Trusted Execution Environments) are only available on commodity (non-specialized) hardware. Machine learning tasks are increasingly executed on specialized hardware accelerators, for which the development of secure enclaves faces significant up-front costs and may not be the most appropriate hardware-based solution. Since AI systems always involve physical infrastructure, the security of that infrastructure can play a key role in claims about a system or its components being secure and private. Secure enclaves have emerged in recent years as a way to demonstrate strong claims about privacy and security that cannot be achieved through software alone. iPhones equipped with facial recognition for screen unlocking, for example, store face-related data on a physically distinct part of the computer known as a secure enclave in order to provide more robust privacy protection. Increasing the range of scenarios in which secure enclaves can be applied in AI, as discussed in this subsection, would enable higher degrees of security and privacy protection to be demonstrated and demanded. A secure enclave is a set of software and hardware features that together provide an isolated execution environment that enables a set of strong guarantees regarding security for applications running inside the enclave [109]. Secure enclaves reduce the ability of malicious actors to access sensitive data or interfere with a program, even if they have managed to gain access to the system outside the enclave. Secure enclaves provide these guarantees by linking high-level desired properties (e.g., isolation of a process from the rest of the system) to low-level design of the chip layout and low-level software interacting with the chip. The connection between physical design and low-level software and high-level security claims relies on a set of underlying assumptions. Despite the fact that researchers have been able to find ways to invalidate these underlying assumptions in some cases, and thus invalidate the high-level security claims [110] [111], these mechanisms help to focus defensive efforts and assure users that relatively extreme measures would be required to invalidate the claims guaranteed by the design of the enclave. While use of secure enclaves has become relatively commonplace in the commodity computing industries, their use in machine learning is less mature. Execution of machine learning on secure enclaves has been demonstrated, but comes with a performance overhead [112]. Demonstrations to date have been carried out on commodity hardware (CPUs [113] [114] and GPUs [115]) or have secure and verifiable outsourcing of parts of the computation to less secure hardware [116] [117], rather than on hardware directly optimized for machine learning (such as TPUs). For most machine learning applications, the cost of using commodity hardware not specialized for machine learning is fairly low because the hardware already exists, and their computational demands can be met on such commodity hardware. However, cutting edge machine learning models often use significantly more computational resources [118], driving the use of more specialized hardware for both training and inference. If used with specialized AI hardware, the use of secure enclaves would require renewed investment for every new design, which can end up being very costly if generations are short and of limited batch sizes (as the cost is amortized across all chips that use the design). Some specialized AI hardware layouts may require entirely novel hardware security features – as the secure enclave model may not be applicable – involving additional costs. One particularly promising guarantee that might be provided by ML-specific hardware security features, coupled with some form of remote attestation, is a guarantee that a model will never leave a particular chip, which could be a key building block of more complex privacy and security policies.

Stakeholder(s):

Industry

Academia

8.1. Specialized Hardware

Integrate security features into ML-specialized hardware.

A focused and ongoing effort to integrate hardware security features into ML-specialized hardware could add value, though it will require collaboration across the sector.

8.2. Enclaves

Open source secure enclave designs.

Recent efforts to open source secure enclave designs could help accelerate the process of comprehensively analyzing the security claims made about certain systems [119]. As more workloads move to specialized hardware, it will be important to either develop secure enclaves for such hardware (or alternative hardware security solutions), or otherwise define best practices for outsourcing computation to "untrusted" accelerators while maintaining privacy and security. Similarly, as many machine learning applications are run on GPUs, it will be important to improve the practicality of secure enclaves or equivalent privacy protections on these processors.

8.3. Other Alternatives

Consider other forms of security features for hardware-based ML accelerators.

The addition of dedicated security features to ML accelerators at the hardware level may need to take a different form than a secure enclave. This is in part due to different architectures and different use of space on the chip; in part due to different weighting of security concerns (e.g., it may be especially important to prevent unauthorized access to user data); and in part due to a difference in economies of scale relative to commodity chips, with many developers of ML accelerators being smaller, less-well-resourced actors relative to established chip design companies like Intel or NVIDIA.

Recommendation 9. Computing Power

Estimate the computing power involved in a single project in great detail.

Problem: The absence of standards for measuring the use of computational resources reduces the value of voluntary reporting and makes it harder to verify claims about the resources used in the AI development process. Although we cannot know for certain due to limited transparency, it is reasonable to assume that a significant majority of contemporary computing hardware used for AI training and inference is installed in data centers (which could be corporate, governmental, or academic), with smaller fractions in server rooms or attached to individual PCs. Many tools and systems already exist to monitor installed hardware and compute usage internally (e.g., across a cloud provider's data center or across an academic cluster's user base). A current example of AI developers reporting on their compute usage is the inclusion of training-related details in published research papers and pre-prints, which often share the amount of compute used to train or run a model. These are done for the purposes of comparison and replication, though often extra work is required to make direct comparisons as there is no standard method for measuring and reporting compute usage. This ambiguity poses a challenge to trustworthy AI development, since even AI developers who want to make verifiable claims about their hardware use are not able to provide such information in a standard form that is comparable across organizations and contexts. Even in the context of a particular research project, issues such as mixed precision training, use of heterogeneous computing resources, and use of pretrained models all complicate accurate reporting that is comparable across organizations. The lack of a common standard or accepted practice on how to report the compute resources used in the context of a particular project has led to several efforts to extract or infer the computational requirements of various advances and compare them using a common framework [118]. The challenge of providing accurate and useful information about the computational requirements of a system or research project is not unique to AI – computer systems research has struggled with this problem for some time. Both fields have seen an increasing challenge in comparing and reproducing results now that organizations with exceptionally large compute resources (also referred to as "hyperscalers") play an ever-increasing role in research in those fields. We believe there is value in further engaging with the computer systems research community to explore challenges of reproducibility, benchmarking, and reporting, though we also see value in developing AI-specific standards for compute reporting. Increasing the precision and standardization of

compute reporting could enable easier comparison of research results across organizations. Improved methods could also serve as building blocks of credible third party oversight of AI projects: an auditor might note, for example, that an organization has more computing power available to it than was reportedly used on an audited project, and thereby surface unreported activities relevant to that project. And employees of an organization are better able to ensure that their organization is acting responsibly to the extent that they are aware of how computing power, data, and personnel are being allocated internally for different purposes.

Stakeholder(s):

AI Labs :
One or more AI labs should estimate the computing power involved in a single project in great detail (high-precision compute measurement), and report on the potential for wider adoption of such methods.

9.1. Standardization

Assess the feasibility of standardizing compute accounting.

We see value in one or more AI labs conducting a "comprehensive" compute accounting effort, as a means of assessing the feasibility of standardizing such accounting. "Comprehensive" here refers to accounting for as much compute usage pertinent to the project as is feasible, and increasing the precision of reported results relative to existing work. It is not clear how viable standardization is, given the aforementioned challenges, though there is likely room for at least incremental progress: just in the past few years, a number of approaches to calculating and reporting compute usage have been tried, and in some cases have propagated across organizations.

Stakeholder(s):

AI Labs

AI Researchers :
AI researchers interested in conducting such a pilot should work with computer systems researchers who have worked on related challenges in other contexts, including the automating of logging and reporting.

Computer Systems Researchers

AI Experts :
Notably, accounting of this sort has costs associated with it, and the metrics of success are unclear. Some accounting efforts could be useful for experts but inaccessible to non-experts, for example, or could only be workable in a particular context (e.g., with a relatively simple training and inference pipeline and limited use of pretrained models). As such, we do not advocate for requiring uniformly comprehensive compute reporting.

9.2. Reporting

Automate or simplify compute reporting.

Depending on the results of early pilots, new tools might help automate or simplify such reporting, though this is uncertain. One reason for optimism about the development of a standardized approach is that a growing fraction of computing power usage occurs in the cloud at "hyperscale" data centers, so a relatively small number of actors could potentially implement best practices that apply to a large fraction of AI development [120].

Stakeholder(s):

Hyperscale Data Centers

AI Researchers :
It is also at present unclear who should have access to reports about compute accounting. While we applaud the current norm in AI research to voluntarily share compute requirements publicly, we expect for-profit entities would have to balance openness with commercial secrecy, and government labs may need to balance openness with security considerations.

Commercial Entities

Government Labs

The Public

Auditors :
This may be another instance in which auditors or independent verifiers could play a role. Standardization of compute accounting is one path to formalizing the auditing practice in this space, potentially as a building block to more holistic auditing regimes. However, as with other mechanisms discussed here, it is insufficient on its own.

Recommendation 10. Funding

Increase funding of computing power resources to improve the ability to verify claims.

Problem: The gap in compute resources between industry and academia limits the ability of those outside of industry to scrutinize technical claims made by AI developers, particularly those related to compute-intensive systems. In recent years, a large number of academic AI researchers have transitioned into industry AI labs. One reason for this shift is the greater availability of computing resources in industry compared to academia. This talent shift has resulted in a range of widely useful software frameworks and algorithmic insights, but has also raised concerns about the growing disparity between the computational resources available to academia and industry [125]. The disparity between industry and academia is clear overall, even though some academic labs are generously supported by government or industry sponsors, and some government agencies are on the cutting edge of building and providing access to supercomputers. Here we focus on a specific benefit of governments taking action to level the playing field of computing power: namely, improving the ability of financially disinterested parties such as academics to verify the claims made by AI developers in industry, especially in the context of compute-intensive systems. Example use cases include:

Stakeholder(s):

Government Funding Bodies :
Government funding bodies should substantially increase funding of computing power resources for researchers in academia, in order to improve the ability of those researchers to verify claims made by industry... While computing power is not a panacea for addressing the gap in resources available for research in academia and industry, funding bodies such as those in governments could level the playing field between sectors by more generously providing computing credits to researchers in academia. Such compute provision could be made more affordable by governments leveraging their purchasing power in negotiations over bulk compute purchases. Governments could also build their own compute infrastructures for this purpose. The particular amounts of compute in question, securing the benefits of scale while avoiding excessive dependence on a particular compute provider, and ways of establishing appropriate terms for the use of such compute are all exciting areas for future research.

Academia

Industry

10.1. Open-Sourcing

Provide open-source alternatives to commercial AI systems.

Providing open-source alternatives to commercial AI systems: given the current norm in AI development of largely-open publication of research, a limiting factor in providing open source alternatives to commercially trained AI models is often the computing resources required. As models become more compute-intensive, government support may be required to maintain a thriving open source AI ecosystem and the various benefits that accrue from it.

10.2. Commercial Models

Increase scrutiny of commercial models.

Increasing scrutiny of commercial models: as outlined in the institutional mechanisms section (see the subsections on red team exercises and bias and safety bounties), there is considerable value in independent third parties stress-testing the models developed by others. While "black box" testing can take place without access to significant compute resources (e.g., by remote access to an instance of the system), local replication for the purpose of testing could make testing easier, and could uncover further issues than those surfaced via remote testing alone. Additional computing resources may be especially needed for local testing of AI systems that are too large to run on a single computer (such as some recent language models).

10.3. AI

Leverage AI to test AI.

Leveraging AI to test AI: as AI systems become more complex, it may be useful or even necessary to deploy adaptive, automated tests to explore potential failure modes or hidden biases, and such testing may become increasingly compute-intensive.

10.4. Claims

Verify claims about compute requirements.

Verifying claims about compute requirements: as described above, accounting for the compute inputs of model training is currently an open challenge in AI development. In tandem with standardization of compute accounting, compute support to non-industry actors would enable replication efforts, which would verify or undermine claims made by AI developers about the resource requirements of the systems they develop.

Administrative Information
Start Date: 2020-04-30
End Date:

Publication Date: 2020-06-15
Source: https://arxiv.org/pdf/2004.07213.pdf

Submitter:
Given Name: Owen
Surname: Ambur
Email: Owen.Ambur@verizon.net
Phone: