Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al.

HAL Id: hal-03065927

https://inria.hal.science/hal-03065927

Preprint submitted on 15 Dec 2020

Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*

Miles Brundage[1†], Shahar Avin[3,2†], Jasmine Wang[4,29†‡] Haydn Belfield[3,2†], Gretchen Krueger[1†],
Gillian Hadfield[1,5,30], Heidy Khlaaf[6], Jingying Yang[7], Helen Toner[8], Ruth Fong[9],
Tegan Maharaj[4,28], Pang Wei Koh[10], Sara Hooker[11], Jade Leung[12], Andrew Trask[9],
Emma Bluemke[9], Jonathan Lebensold[4,29], Cullen O'Keefe[1], Mark Koren[13], Théo Ryffel[14],
JB Rubinovitz[15], Tamay Besiroglu[16], Federica Carugati[17], Jack Clark[1], Peter Eckersley[7],
Sarah de Haas[18], Maritza Johnson[18], Ben Laurie[18], Alex Ingerman[18], Igor Krawczuk[19],
Amanda Askell[1], Rosario Cammarota[20], Andrew Lohn[21], David Krueger[4,27], Charlotte Stix[22],
Peter Henderson[10], Logan Graham[9], Carina Prunkl[12], Bianca Martin[1], Elizabeth Seger[16],
Noa Zilberman[9], Seán Ó hÉigeartaigh[2,3], Frens Kroeger[23], Girish Sastry[1], Rebecca Kagan[8],
Adrian Weller[16,24], Brian Tse[12,7], Elizabeth Barnes[1], Allan Dafoe[12,9], Paul Scharre[25],
Ariel Herbert-Voss[1], Martijn Rasser[25], Shagun Sodhani[4,27], Carrick Flynn[8],
Thomas Krendl Gilbert[26], Lisa Dyer[7], Saif Khan[8], Yoshua Bengio[4,27], Markus Anderljung[12]

[1]OpenAI, [2]Leverhulme Centre for the Future of Intelligence, [3]Centre for the Study of Existential Risk,
[4]Mila, [5]University of Toronto, [6]Adelard, [7]Partnership on AI, [8]Center for Security and Emerging Technology,
[9]University of Oxford, [10]Stanford University, [11]Google Brain, [12]Future of Humanity Institute,
[13]Stanford Centre for AI Safety, [14]École Normale Supérieure (Paris), [15]Remedy.AI,
[16]University of Cambridge, [17]Center for Advanced Study in the Behavioral Sciences, [18]Google Research,
[19]École Polytechnique Fédérale de Lausanne, [20]Intel, [21]RAND Corporation,
[22]Eindhoven University of Technology, [23]Coventry University, [24]Alan Turing Institute,
[25]Center for a New American Security, [26]University of California, Berkeley,
[27]University of Montreal, [28]Montreal Polytechnic, [29]McGill University,
[30]Schwartz Reisman Institute for Technology and Society

April 2020

Contents

Executive Summary

Recent progress in artificial intelligence (AI) has enabled a diverse array of applications across commercial, scientific, and creative domains. With this wave of applications has come a growing awareness of the large-scale impacts of AI systems, and recognition that existing regulations and norms in industry and academia are insufficient to ensure responsible AI development [1] [2] [3].

Steps have been taken by the AI community to acknowledge and address this insufficiency, including widespread adoption of ethics principles by researchers and technology companies. However, ethics principles are non-binding, and their translation to actions is often not obvious. Furthermore, those outside a given organization are often ill-equipped to assess whether an AI developer's actions are consistent with their stated principles. Nor are they able to hold developers to account when principles and behavior diverge, fueling accusations of "ethics washing" [4]. In order for AI developers to earn trust from system users, customers, civil society, governments, and other stakeholders that they are building AI responsibly, there is a need to move beyond principles to a focus on mechanisms for demonstrating responsible behavior [5]. Making and assessing verifiable claims, to which developers can be held accountable, is one crucial step in this direction.

With the ability to make precise claims for which evidence can be brought to bear, AI developers can more readily demonstrate responsible behavior to regulators, the public, and one another. Greater verifiability of claims about AI development would help enable more effective oversight and reduce pressure to cut corners for the sake of gaining a competitive edge [1]. Conversely, without the capacity to verify claims made by AI developers, those using or affected by AI systems are more likely to be put at risk by potentially ambiguous, misleading, or false claims.

This report suggests various steps that different stakeholders in AI development can take to make it easier to verify claims about AI development, with a focus on providing evidence about the safety, security, fairness, and privacy protection of AI systems. Implementation of such mechanisms can help make progress on the multifaceted problem of ensuring that AI development is conducted in a trustworthy fashion.[1] The mechanisms outlined in this report deal with questions that various parties involved in AI development might face, such as:

- Can I (as a user) verify the claims made about the level of privacy protection guaranteed by a new AI system I'd like to use for machine translation of sensitive documents?

- Can I (as a regulator) trace the steps that led to an accident caused by an autonomous vehicle? Against what standards should an autonomous vehicle company's safety claims be compared?

- Can I (as an academic) conduct impartial research on the impacts associated with large-scale AI systems when I lack the computing resources of industry?

- Can I (as an AI developer) verify that my competitors in a given area of AI development will follow best practices rather than cut corners to gain an advantage?

Even AI developers who have the desire and/or incentives to make concrete, verifiable claims may not be equipped with the appropriate mechanisms to do so. The AI development community needs a robust "toolbox" of mechanisms to support the verification of claims about AI systems and development processes.

---

[1]The capacity to verify claims made by developers, on its own, would be insufficient to ensure responsible AI development. Not all important claims admit verification, and there is also a need for oversight agencies such as governments and standards organizations to align developers' incentives with the public interest.

This problem framing led some of the authors of this report to hold a workshop in April 2019, aimed at expanding the toolbox of mechanisms for making and assessing verifiable claims about AI development.[2] This report builds on the ideas proposed at that workshop. The mechanisms outlined do two things:

- They increase the options available to AI developers for substantiating claims they make about AI systems' properties.

- They increase the specificity and diversity of demands that can be made of AI developers by other stakeholders such as users, policymakers, and members of civil society.

Each mechanism and associated recommendation discussed in this report addresses a specific gap preventing effective assessment of developers' claims today. Some of these mechanisms exist and need to be extended or scaled up in some way, and others are novel. The report is intended as an incremental step toward improving the verifiability of claims about AI development.

The report organizes mechanisms under the headings of *Institutions*, *Software*, and *Hardware*, which are three intertwined components of AI systems and development processes.

- **Institutional Mechanisms**: These mechanisms shape or clarify the incentives of people involved in AI development and provide greater visibility into their behavior, including their efforts to ensure that AI systems are safe, secure, fair, and privacy-preserving. Institutional mechanisms play a foundational role in verifiable claims about AI development, since it is *people* who are ultimately responsible for AI development. We focus on **third party auditing**, to create a robust alternative to self-assessment of claims; **red teaming exercises**, to demonstrate AI developers' attention to the ways in which their systems could be misused; **bias and safety bounties**, to strengthen incentives to discover and report flaws in AI systems; and **sharing of AI incidents**, to improve societal understanding of how AI systems can behave in unexpected or undesired ways.

- **Software Mechanisms**: These mechanisms enable greater understanding and oversight of specific AI systems' properties. We focus on **audit trails**, to enable accountability for high-stakes AI systems by capturing critical information about the development and deployment process; **interpretability**, to foster understanding and scrutiny of AI systems' characteristics; and **privacy-preserving machine learning**, to make developers' commitments to privacy protection more robust.

- **Hardware Mechanisms**: Mechanisms related to computing hardware can play a key role in substantiating strong claims about privacy and security, enabling transparency about how an organization's resources are put to use, and influencing who has the resources necessary to verify different claims. We focus on **secure hardware for machine learning**, to increase the verifiability of privacy and security claims; **high-precision compute measurement**, to improve the value and comparability of claims about computing power usage; and **compute support for academia**, to improve the ability of those outside of industry to evaluate claims about large-scale AI systems.

Each mechanism provides additional paths to verifying AI developers' commitments to responsible AI development, and has the potential to contribute to a more trustworthy AI ecosystem. The full list of recommendations associated with each mechanism is found on the following page and again at the end of the report.

---

[2] See **Appendix I, "Workshop and Report Writing Process."**

List of Recommendations

Institutional Mechanisms and Recommendations

1. A coalition of stakeholders should create a task force to research options for conducting and funding **third party auditing** of AI systems.

2. 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.

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

4. AI developers should share more information about **AI incidents**, including through collaborative channels.

Software Mechanisms and Recommendations

5. Standards setting bodies should work with academia and industry to develop **audit trail** requirements for safety-critical applications of AI systems.

6. Organizations developing AI and funding bodies should support research into the **interpretability** of AI systems, with a focus on supporting risk assessment and auditing.

7. AI developers should develop, share, and use suites of tools for **privacy-preserving machine learning** that include measures of performance against common standards.

Hardware Mechanisms and Recommendations

8. 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.

9. 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.

10. 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.

1 Introduction

1.1 Motivation

With rapid technical progress in artificial intelligence (AI)[3] and the spread of AI-based applications over the past several years, there is growing concern about how to ensure that the development and deployment of AI is beneficial – and not detrimental – to humanity. In recent years, AI systems have been developed in ways that are inconsistent with the stated values of those developing them. This has led to a rise in concern, research, and activism relating to the impacts of AI systems [2] [3]. AI development has raised concerns about amplification of bias [6], loss of privacy [7], digital addictions [8], social harms associated with facial recognition and criminal risk assessment [9], disinformation [10], and harmful changes to the quality [11] and availability of gainful employment [12].

In response to these concerns, a range of stakeholders, including those developing AI systems, have articulated ethics principles to guide responsible AI development. The amount of work undertaken to articulate and debate such principles is encouraging, as is the convergence of many such principles on a set of widely-shared concerns such as safety, security, fairness, and privacy.[4]

However, **principles are only a first step in the effort to ensure beneficial societal outcomes from AI** [13]. Indeed, studies [17], surveys [18], and trends in worker and community organizing [2] [3] make clear that large swaths of the public are concerned about the risks of AI development, and do not trust the organizations currently dominating such development to self-regulate effectively. Those potentially affected by AI systems need mechanisms for ensuring responsible development that are more robust than high-level principles. People who get on airplanes don't trust an airline manufacturer because of its PR campaigns about the importance of safety - they trust it because of the accompanying infrastructure of technologies, norms, laws, and institutions for ensuring airline safety.[5] Similarly, along with the growing explicit adoption of ethics principles to guide AI development, there is mounting skepticism about whether these claims and commitments can be monitored and enforced [19].

Policymakers are beginning to enact regulations that more directly constrain AI developers' behavior [20]. We believe that analyzing AI development through the lens of verifiable claims can help to inform such efforts. AI developers, regulators, and other actors all need to understand which properties of AI systems and development processes can be credibly demonstrated, through what means, and with what tradeoffs.

We define **verifiable claims**[6] as falsifiable statements for which evidence and arguments can be brought

---

[3]We define AI as digital systems that are capable of performing tasks commonly thought to require intelligence, with these tasks typically learned via data and/or experience.

[4]Note, however, that many such principles have been articulated by Western academics and technology company employees, and as such are not necessarily representative of humanity's interests or values as a whole. Further, they are amenable to various interpretations [13][14] and agreement on them can mask deeper disagreements [5]. See also Beijing AI Principles [15] and Zeng et. al. [16] for examples of non-Western AI principles.

[5]Recent commercial airline crashes also serve as a reminder that even seemingly robust versions of such infrastructure are imperfect and in need of constant vigilance.

[6]While this report does discuss the technical area of formal verification at several points, and several of our recommendations are based on best practices from the field of information security, the sense in which we use "verifiable" is distinct from how the term is used in those contexts. Unless otherwise specified by the use of the adjective "formal" or other context, this report uses the word verification in a looser sense. Formal verification seeks mathematical proof that a certain technical claim is

to bear on the likelihood of those claims being true. While the degree of attainable certainty will vary across different claims and contexts, we hope to show that greater degrees of evidence can be provided for claims about AI development than is typical today. The nature and importance of verifiable claims is discussed in greater depth in **Appendix III**, and we turn next to considering the types of mechanisms that can make claims verifiable.

1.2   Institutional, Software, and Hardware Mechanisms

AI developers today have many possible approaches for increasing the verifiability of their claims. Despite the availability of many mechanisms that could help AI developers demonstrate their claims and help other stakeholders scrutinize their claims, this toolbox has not been well articulated to date.

We view AI development processes as sociotechnical systems,[7] with institutions, software, and hardware all potentially supporting (or detracting from) the verifiability of claims about AI development. AI developers can make claims about, or take actions related to, each of these three interrelated pillars of AI development.

In some cases, adopting one of these mechanisms can increase the verifiability of one's own claims, whereas in other cases the impact on trust is more indirect (i.e., a mechanism implemented by one actor enabling greater scrutiny of other actors). As such, collaboration across sectors and organizations will be critical in order to build an ecosystem in which claims about responsible AI development can be verified.

- Institutional mechanisms largely pertain to *values, incentives, and accountability*. Institutional mechanisms shape or clarify the incentives of people involved in AI development and provide greater visibility into their behavior, including their efforts to ensure that AI systems are safe, secure, fair, and privacy-preserving. These mechanisms can also create or strengthen channels for holding AI developers accountable for harms associated with AI development. In this report, we provide an overview of some such mechanisms, and then discuss **third party auditing**, **red team exercises**, **safety and bias bounties**, and **sharing of AI incidents** in more detail.

- Software mechanisms largely pertain to *specific AI systems and their properties*. Software mechanisms can be used to provide evidence for both formal and informal claims regarding the properties of specific AI systems, enabling greater understanding and oversight. The software mechanisms we highlight below are **audit trails**, **interpretability**, and **privacy-preserving machine learning**.

- Hardware mechanisms largely pertain to *physical computational resources and their properties*. Hardware mechanisms can support verifiable claims by providing greater assurance regarding the privacy and security of AI systems, and can be used to substantiate claims about how an organization is using their general-purpose computing capabilities. Further, the distribution of resources across different actors can influence the types of AI systems that are developed and which actors are capable of assessing other actors' claims (including by reproducing them). The hardware mechanisms we focus on in this report are **hardware security features for machine learning**, **high-precision compute measurement**, and **computing power support for academia**.

---

true with certainty (subject to certain assumptions). In contrast, this report largely focuses on claims that are unlikely to be demonstrated with absolute certainty, but which can be shown to be likely or unlikely to be true through relevant arguments and evidence.

[7]Broadly, a sociotechnical system is one whose "core interface consists of the relations between a nonhuman system and a human system", rather than the components of those systems in isolation. See Trist [21].

1.3   Scope and Limitations

This report focuses on a particular aspect of trustworthy AI development: the extent to which organizations developing AI systems can and do make verifiable claims about the AI systems they build, and the ability of other parties to assess those claims. Given the backgrounds of the authors, the report focuses in particular on mechanisms for demonstrating claims about AI systems being safe, secure, fair, and/or privacy-preserving, without implying that those are the only sorts of claims that need to be verified.

We devote particular attention to mechanisms[8] that the authors have expertise in and for which concrete and beneficial next steps were identified at an April 2019 workshop. These are not the only mechanisms relevant to verifiable claims; we survey some others at the beginning of each section, and expect that further useful mechanisms have yet to be identified.

Making verifiable claims is part of, but not equivalent to, trustworthy AI development, broadly defined. An AI developer might also be more or less trustworthy based on the particular values they espouse, the extent to which they engage affected communities in their decision-making, or the extent of recourse that they provide to external parties who are affected by their actions. Additionally, the actions of AI developers, which we focus on, are not all that matters for trustworthy AI development–the existence and enforcement of relevant laws matters greatly, for example.

Appendix I discusses the reasons for the report's scope in more detail, and **Appendix II** discusses the relationship between different definitions of trust and verifiable claims. When we use the term "trust" as a verb in the report, we mean that one party (party A) gains confidence in the reliability of another party's claims (party B) based on evidence provided about the accuracy of those claims or related ones. We also make reference to this claim-oriented sense of trust when we discuss actors "earning" trust, (providing evidence for claims made), or being "trustworthy" (routinely providing sufficient evidence for claims made). This use of language is intended to concisely reference an important dimension of trustworthy AI development, and is not meant to imply that verifiable claims are sufficient for attaining trustworthy AI development.

1.4   Outline of the Report

The next three sections of the report, **Institutional Mechanisms and Recommendations**, **Software Mechanisms and Recommendations**, and **Hardware Mechanisms and Recommendations**, each begin with a survey of mechanisms relevant to that category. Each section then highlights several mechanisms that we consider especially promising. We are uncertain which claims are most important to verify in the context of AI development, but strongly suspect that some combination of the mechanisms we outline in this report are needed to craft an AI ecosystem in which responsible AI development can flourish.

The way we articulate the case for each mechanism is *problem-centric*: each mechanism helps address a potential barrier to claim verification identified by the authors. Depending on the case, the recommendations associated with each mechanism are aimed at implementing a mechanism for the first time, researching it, scaling it up, or extending it in some way.

---

[8]We use the term mechanism generically to refer to processes, systems, or approaches for providing or generating evidence about behavior.

The **Conclusion** puts the report in context, discusses some important caveats, and reflects on next steps.

The **Appendices** provide important context, supporting material, and supplemental analysis. **Appendix I** provides background on the workshop and the process that went into writing the report; **Appendix II** serves as a glossary and discussion of key terms used in the report; **Appendix III** discusses the nature and importance of verifiable claims; **Appendix IV** discusses the importance of verifiable claims in the context of arms control; **Appendix V** provides context on antitrust law as it relates to cooperation among AI developers on responsible AI development; and **Appendix VI** offers supplemental analysis of several mechanisms.

2   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[9] 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.

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.[10]

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.

---

[9]**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.

[10]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 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].

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.[11]

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[12] 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.

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].[13]

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;[14] 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

---

[11]The Windfall Clause [30] is one proposal along these lines, and involves an ex ante commitment by AI firms to donate a significant amount of any eventual extremely large profits.

[12]The global nature of AI development, and the national nature of much relevant regulation, is a key complicating factor.

[13]See for example the norms set and enforced by the European Telecommunications Standards Institute (ETSI). These norms have real "teeth," such as the obligation for designated holders of Standard Essential Patents to license on Fair, Reasonable and Non-discriminatory (FRAND) terms. Breach of FRAND could give rise to a breach of contract claim as well as constitute a breach of antitrust law [32]. Voluntary standards for consumer products, such as those associated with Fairtrade and Organic labels, are also potentially relevant precedents [33].

[14]An example of such an effort is the Asilomar AI Principles [34].

for maintaining and deepening interorganizational communication [10].[15]

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.[16] 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.

The subsections below each introduce and explore a mechanism with the potential for improving the verifiability of claims in AI development: **third party auditing**, **red team exercises**, **bias and safety bounties**, and **sharing of AI incidents**. In each case, the subsections below begin by discussing a problem which motivates exploration of that mechanism, followed by a recommendation for improving or applying that mechanism.

---

[15]Though note competitors sharing commercially sensitive, non-public information (such as strategic plans or R&D plans) could raise antitrust concerns. It is therefore important to have the right antitrust governance structures and procedures in place (i.e., setting out exactly what can and cannot be shared). See **Appendix V**.

[16]Recent revelations regarding the culture of engineering and management at Boeing highlight the urgency of this issue [35].

2.1 Third Party Auditing

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.

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.

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.[17] 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].

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.

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.[18] 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.[19]

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-

---

[17]See Raji and Smart et al. [36] for a discussion of some lessons for AI from auditing in other industries.

[18]Model cards for model reporting [28] and data sheets for datasets [29] reveal information about AI systems publicly, and future work in third party auditing could build on such tools, as advocated by Raji and Smart et al. [36].

[19]Consumer Reports, originally founded as the Consumers Union in 1936, is one model for an independent, third party organization that performs similar functions for products that can affect the health, well-being, and safety of the people using those products. (https://www.consumerreports.org/cro/about-us/what-we-do/research-and-testing/index.htm).

[20]See **Appendix III** for further discussion of claim-based frameworks for auditing.

evolving societal challenges posed by AI.

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

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]. 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).

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.

Recommendation: A coalition of stakeholders should create a task force to research options for conducting and funding third party auditing of AI systems.

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.[21] 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.

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.[22] Regardless of whether such an approach is pursued, third party auditing by private actors should be viewed as a complement

---

[21]This list is not exhaustive - see, e.g., [39], [40], and [41] for related discussions.
[22]Examples of such entities include EXIDA, the UK Office of Nuclear Regulation, and the private company Adelard.

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.

2.2   Red Team Exercises

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.

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.[23]

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.

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.[24] 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 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.[25] 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.

Doing red teaming in a more collaborative fashion, as a community of focused professionals across

---

[23]Red teaming could be aimed at assessing various properties of AI systems, though we focus on safety and security in this subsection given the expertise of the authors who contributed to it.

[24]For an example of early efforts related to this, see Marshall et al., "Threat Modeling AI/ML Systems and Dependencies" [43]

[25]In the context of language models, for example, 2019 saw a degree of communication and coordination across AI developers to assess the relative risks of different language understanding and generation systems [10]. Adversarial machine learning, too, is an area with substantial sharing of lessons across organizations, though it is not obvious whether a shared red team focused on this would be too broad.

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.[26]

Recommendation: 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.

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?[27] 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.

---

[26]This has a precedent from cybersecurity; MITRE's ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations, which serves as a foundation for development of more specific threat models and methodologies to improve cybersecurity (https://attack.mitre.org/).

[27]These practical questions are not exhaustive, and even addressing them effectively might not suffice to ensure that collaborative red teaming is beneficial. For example, one potential failure mode is if collaborative red teaming fostered excessive homogeneity in the red teaming approaches used, contributing to a false sense of security in cases where that approach is insufficient.

2.3 Bias and Safety Bounties

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.

"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.

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.

Bias[28] 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.[29] 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.[30]

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]. 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]. To date, investigative journalists and civil society organizations have played key roles in surfacing different biases in deployed AI systems. 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.[31]

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,[32] though in both

---

[28]For an earlier exploration of bias bounties by one of the report authors, see Rubinovitz [46].

[29]For example, model cards for model reporting [28] and datasheets for datasets [29] are recently developed means of documenting AI releases, and such documentation could be extended with publicly listed incentives for finding new forms of problematic behavior not captured in that documentation.

[30]Bounties for finding issues with datasets used for training AI systems could also be considered, though we focus on trained AI systems and code as starting points.

[31]We note that many millions of dollars have been paid to date via bug bounty programs in the computer security domain, providing some evidence for this hypothesis. However, bug bounties are not a panacea and recourse to the public is also appropriate in some cases.

[32]We also note that the challenge of avoiding harmful biases is sometimes framed as a subset of safety, though for the

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.

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.

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

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;[33]

- 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,[34] 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

---

purposes of this discussion, little hinges on this terminological issue. We distinguish the two in the title of this section in order to call attention to the unique properties of different types of bounties.

[33]Note that we specifically consider public bounty programs here, though instances of private bounty programs also exist in the computer security community. Even in the event of a publicly advertised bounty, however, submissions may be private, and as such there is a need for explicit policies for handling submissions in a timely and legitimate fashion–otherwise such programs will provide little assurance.

[34]A zero-day vulnerability is a security vulnerability that is unknown to the developers of the system and other affected parties, giving them "zero days" to mitigate the issue if the vulnerability were to immediately become widely known. The computer security community features a range of views on appropriate responses to zero-days, with a common approach being to provide a finite period for the vendor to respond to notification of the vulnerability before the discoverer goes public.

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.

2.4 Sharing of 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.**

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 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.

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.[35] The pilot was built using publicly available information through a set of volunteers and contractors manually collecting known AI incidents where AI caused harm in the real world.

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;

---

[35] See Partnership on AI's AI Incident Registry as an example (http://aiid.partnershiponai.org/). A related resource is a list called Awful AI, which is intended to raise awareness of misuses of AI and to spur discussion around contestational research and tech projects [51]. A separate list summarizes various cases in which AI systems "gamed" their specifications in unexpected ways [52]. Additionally, AI developers have in some cases provided retrospective analyses of particular AI incidents, such as with Microsoft's "Tay" chatbot [53].

- 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.[36]

Recommendation: 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.

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).

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.

---

[36]We do not mean to claim that building and using such infrastructure would be sufficient to ensure that AI incidents are addressed effectively. Sharing is only one part of the puzzle for effectively managing incidents. For example, attention should also be paid to ways in which organizations developing AI, and particularly safety-critical AI, can become "high reliability organizations" (see, e.g., [54]).

3 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.[37] 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.[38]

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[39]; and

- This system provides repeatable or reproducible results.

Below, we summarize several clusters of mechanisms which help to substantiate some of the claims above.

Reproducibility of technical results in AI is a key way of enabling verification of claims about system

---

[37]For example, consider a desideratum for privacy: access to a dataset should not enable an adversary to learn anything about an individual that could not be learned without access to the database. Differential privacy as originally conceived does not guarantee this–rather, it guarantees (to an extent determined by a privacy budget) that one cannot learn whether that individual was in the database in question.

[38]In **Section 3.3**, we discuss the role that computing power–in addition to expertise–can play in influencing who can verify which claims.

[39]Conceptions of, and measures for, fairness in machine learning, philosophy, law, and beyond vary widely. See, e.g., Xiang and Raji [59] and Binns [60].

properties, and a number of ongoing initiatives are aimed at improving reproducibility in AI.[40][41] 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.[42]

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)[43] 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,[44] which we discuss in **Appendix VI(A)**.

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 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

---

[40]We note the distinction between narrow senses of reproducibility that focus on discrete technical results being reproducible given the same initial conditions, sometimes referred to as repeatability, and broader senses of reproducibility that involve reported performance gains carrying over to different contexts and implementations.

[41]One way to promote robustness is through incentivizing reproducibility of reported results. There are increasing effort to award systems the recognition that they are robust, e.g., through ACM's artifact evaluation badges https://www.acm.org/publications/policies/artifact-review-badging. Conferences are also introducing artifact evaluation, e.g., in the intersection between computer systems research and ML. See, e.g., https://reproindex.com/event/repro-sml2020 and http://cknowledge.org/request.html The Reproducibility Challenge is another notable effort in this area: https://reproducibility-challenge.github.io/neurips2019/

[42]In the following section on hardware mechanisms, we also discuss how reproducibility can be advanced in part by leveling the playing field between industry and other sectors with respect to computing power.

[43]Machine learning is a subfield of AI focused on the design of software that improves in response to data, with that data taking the form of unlabeled data, labeled data, or experience. While other forms of AI that do not involve machine learning can still raise privacy concerns, we focus on machine learning here given the recent growth in associated privacy techniques as well as the widespread deployment of machine learning.

[44]Research into perception-based properties such as pointwise robustness, for example, are not sufficiently comprehensive to be applied to real-time critical AI systems such as autonomous vehicles.

models could be robustly attributed via randomized controlled trials.

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.

The subsections below discuss software mechanisms that we consider especially important to advance further. In particular, we discuss **audit trails**, **interpretability**, and **privacy-preserving machine learning**.

---

[45]Adversarial robustness refers to an AI system's ability to perform well in the context of (i.e. to be robust against) "adversarial" inputs, or inputs designed specifically to degrade the system's performance.

3.1 Audit Trails

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,[46] 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.

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.[47] 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: 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.[48] Although application dependent, software audit trails often require a base set of traceability[49] trails to be demonstrated for qualification;[50] the decision to choose a certain set of trails requires considering trade-offs about efficiency, completeness, tamperproofing, and other design considerations. There is flexibility in the type of documents or evidence the auditee presents to satisfy these general traceability requirements

---

[46]Such as the National Transportation Safety Board with regards to autonomous vehicle traffic accidents.

[47]See **Appendix III** for discussion of assurance cases and related frameworks.

[48]Others have argued for the importance of audit trails for AI elsewhere, sometimes under the banner of "logging." See, e.g., [72].

[49]Traceability in this context refers to "the ability to verify the history, location, or application of an item by means of documented recorded identification," https://en.wikipedia.org/wiki/Traceability, where the item in question is digital in nature, and might relate to various aspects of an AI system's development and deployment process.

[50]This includes traceability: between the system safety requirements and the software safety requirements, between the software safety requirements specification and software architecture, between the software safety requirements specification and software design, between the software design specification and the module and integration test specifications, between the system and software design requirements for hardware/software integration and the hardware/software integration test specifications, between the software safety requirements specification and the software safety validation plan, and between the software design specification and the software verification (including data verification) plan.

(e.g., between test logs and requirement documents, verification and validation activities, etc.).[51]

Existing standards often define in detail the required audit trails for specific applications. 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 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.

---

[51]See **Appendix III**.

3.2 Interpretability

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.[52]

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,[53] 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].[54]

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: <u>a.</u> The method should provide sufficient insight for

---

[52]While definitions in this area are contested, some would distinguish between "interpretability" and "explainability" as categories for these two directions, respectively.

[53]See, e.g., Sendak et. al. [76] which focuses on building trust in a hospital context, and contextualizes the role of interpretability in this process.

[54]In New York, an investigation is ongoing into apparent gender discrimination associated with the Apple Card's credit line allowances. This case illustrates the interplay of (a lack of) interpretability and the potential harms associated with automated decision-making systems [79].

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.

Recommendation: Organizations developing AI and funding bodies should support research into the interpretability of AI systems, with a focus on supporting 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 well-developed 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:

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

- 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

- 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)**.

3.3 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

28

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.[55] 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.[56] 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.[57]

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,[58] 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

---

[55]To illustrate how statistical noise can be helpful in protecting privacy, consider the difference between a survey that solicits and retains "raw" answers from individuals, on the one hand, and another survey in which the respondents are asked to flip a coin in order to determine whether they will either provide the honest answer right away or flip the coin again in order to determine which answer to provide. The latter approach would enable individual survey respondents to have plausible deniability regarding their true answers, but those conducting the survey could still learn useful information from the responses, since the noise would largely cancel out at scale. For an accessible discussion of the ideas behind differential privacy and its applications, from which this short summary was adapted, see [99].

[56]For an example of the combination of federated learning and differential privacy, see McMahan et. al. [97].

[57]See `https://www.openmined.com`.

[58]See the **Appendix VI(B)** for further discussion of this project.

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.

Recommendation: AI developers should develop, share, and use suites of tools for privacy-preserving machine learning that include measures of performance against common standards.

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.

Open-source communities projects or projects backed by a particular company can sometimes suffer from a lack of stable funding support[59] 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.

Benchmarks for PPML could help unify goals and measure progress across different groups.[60] 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.[61]

---

[59]Novel approaches to funding open source work should also be considered in this context, such as GitHub's "sponsors" initiative. `https://help.github.com/en/github/supporting-the-open-source-community-with-github-sponsors/about-github-sponsors`

[60]The use of standard tools, guides, and benchmarks can also potentially advance research in other areas, but we focus on privacy-preserving ML in particular here given the backgrounds of the authors who contributed to this subsection. Additionally, we note that some benchmarks have been proposed in the PPML literature for specific subsets of techniques, such as DPComp for differential privacy, but we expect that further exploration of benchmarks across the full spectra of PPML techniques would be valuable.

[61]On the other hand, we note that benchmarks also have potential disadvantages, as they incentivize developers to perform well on the specific benchmark, rather than focusing on the specifics of the intended use case of their product or service, which may significantly diverge from the benchmark setting; the design of benchmarks, and more diverse and adaptive evaluation and comparison methods, is its own technical challenge, as well as an institutional challenge to incentivize appropriate curation and use of benchmarks to establish a common understanding of what is achievable.

4 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].[62] 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.[63] 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).

Existing mechanisms performing one or more of these functions are discussed below.

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,[65] 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).

---

[62]Experts disagree on the extent to which large amounts of computing power are key to progress in AI development. See, e.g., Sutton [104] and Brooks [105] for different opinions about the importance of computing power relative to other factors.

[63]Since training AI systems is more compute-intensive[64] than running them, it is not clear that equivalent computational resources will always be required on the part of those verifying claims about an AI system. However, AI systems are also beginning to require non-trivial computing resources to run, sometimes requiring the model to be split over multiple machines. Additionally, one might need to run an AI system many times in order to verify claims about its characteristics, even if each run is inexpensive. We thus make the conservative assumption that more computing resources would be (all things being equal) helpful to the scrutiny of claims about large-scale AI systems, as discussed below in the context of academic access to computing resources, while recognizing that this may not always be true in particular cases.

[65]http://multicore.doc.ic.ac.uk/tools/GPUVerify/

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].

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. 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.[66]

We now turn to more detailed discussions of three hardware mechanisms that could improve the verifiability of claims: we highlight the importance of **secure hardware for machine learning**, **high-precision compute measurement**, and **computing power support for academia**.

4.1 Secure Hardware for Machine Learning

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].[67] 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

---

[67]For example, training ResNet-32 using Myelin (which utilises a CPU secure enclave) requires 12.9 mins/epoch and results in a final accuracy of 90.8%, whereas the same training on a non-private CPU requires 12.3 mins/epoch and results in a final accuracy of 92.4% [112].

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.

Recommendation: Industry and academia should work together to develop hardware security features for AI accelerators[68] or otherwise establish best practices for the use of secure hardware (including secure enclaves on commodity hardware) in machine learning contexts.

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.

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.

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.

---

[68]An AI accelerator is a form of computing hardware that is specialized to perform an AI-related computation efficiently, rather than to perform general purpose computation.

4.2 High-Precision Compute Measurement

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.[69]

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.[70] 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.[71] 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,[72] use of heterogeneous computing resources, and use of pretrained models all complicate accurate reporting that is comparable across organizations.[73] 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.[74] 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")

---

[69]For reference, the Cisco Global Cloud Index forecasts that the ratio of data center traffic to non-data center traffic by 2021 will be 103:1. When looking just at data centers, they forecast that by 2021, 94% of workloads and compute instances will be processed by cloud data centers, with the remaining 6% processed by traditional data centers. Note, however, that these are for general workloads, not AI specific [120].

[70]Schwartz and Dodge et al. have recommended that researchers always publish financial and computational costs alongside performance increases [121].

[71]There are, however, emerging efforts at standardization in particular contexts. The Transaction Processing Performance Council has a related working group, and efforts like MLPerf are contributing to standardization of some inference-related calculations, though accounting for training remains especially problematic.

[72]Mixed precision refers to the growing use of different binary representations of floating point numbers with varying levels of precision (e.g., 8 bit, 16 bit, 32 bit or 64 bit) at different points of a computation, often trading-off lower precision for higher throughput or performance in ML contexts relative to non-ML contexts. Since an 8 bit floating point operation, say, differs in hardware requirements from a 64 bit floating point operation, traditional measures in terms of Floating Point Operations Per Second (FLOPS) fail to capture this heterogeneity.

[73]For an illustrative discussion of the challenges associated with reporting compute usage for a large-scale AI project, see, e.g., OpenAI's Dota 2 project [122].

[74]See, e.g., Vitek & Kalibera [123] and Hoefler & Belli [124].

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.

Recommendation: One or more AI labs should estimate the computing power involved in a single project in great detail, and report on the potential for wider adoption of such methods.

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. 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.

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.

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].

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. 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.

4.3 Compute Support for Academia

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[75] or industry[76] sponsors, and some government agencies are on the cutting edge of building and providing access to supercomputers.[77]

Here we focus on a specific benefit of governments[78] 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:

- **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.

- **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).

- **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.

---

[75]https://epsrc.ukri.org/research/facilities/hpc/

[76]See, e.g., https://www.tensorflow.org/tfrc, https://aws.amazon.com/grants/ and https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/ for examples of industry support for academic computing.

[77]For example, the US Department of Energy's supercomputing division currently hosts the fastest supercomputer worldwide.

[78]While industry actors can and do provide computing power support to non-industry actors in beneficial ways, the scale and other properties of such programs are likely to be affected by the rises and falls of particular companies' commercial fortunes, and thus are not a reliable long-term solution to the issues discussed here.

- **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.

Recommendation: 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.[79] 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.[80]

---

[79]As advocated by various authors, e.g., Sastry et al. [118], Rasser & Lambert et. al. [126], and Etchemendy and Li [127].

[80]This may require significant levels of funding, and so the benefits should be balanced against the opportunity cost of public spending.

5   Conclusion

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

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. 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.[81]

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.

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.

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].

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

---

[81]The landing page for this report, www.towardtrustworthyai.com, will also be used to share relevant updates after the report's publication.

available to consumers that are consistent with their stated values.

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.

Acknowledgements

References

[1] Amanda Askell, Miles Brundage, and Gillian Hadfield. "The Role of Cooperation in Responsible AI Development". In: *arXiv* (July 2019). arXiv: `1907.04534`. URL: `http://arxiv.org/abs/1907.04534`.

[2] Meredith Whittaker et al. *AI Now Report 2018*. Tech. rep. New York, NY, USA: AI Now Institute, 2018, p. 68. URL: `https://ainowinstitute.org/AI%7B%5C_%7DNow%7B%5C_%7D2018%7B%5C_%7DReport.pdf`.

[3] Kate Crawford et al. *AI Now 2019 Report*. Tech. rep. New York, NY, USA: AI Now Institute, 2019, p. 100. URL: `https://ainowinstitute.org/AI%7B%5C_%7DNow%7B%5C_%7D2019%7B%5C_%7DReport.pdf`.

[4] Elizabeth Gibney. "The battle for ethical AI at the world's biggest machine-learning conference". In: *Nature* (2020). ISSN: 0028-0836. DOI: `10.1038/d41586-020-00160-y`. URL: `http://www.nature.com/articles/d41586-020-00160-y`.

[5] Brent Mittelstadt. "AI Ethics - Too Principled to Fail?" In: *Nature Machine Intelligence* (June 2019). DOI: `10.2139/ssrn.3391293`. arXiv: `1906.06668`. URL: `https://arxiv.org/ftp/arxiv/papers/1906/1906.06668.pdf`.

[6] Karen Hao. *This is how AI bias really happens-and why it's so hard to fix - MIT Technology Review*. 2019. URL: `https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/` (visited on 02/01/2020).

[7] University of California - Berkeley. *Artificial intelligence advances threaten privacy of health data: Study finds current laws and regulations do not safeguard individuals' confidential health information – ScienceDaily*. 2019. URL: `https://www.sciencedaily.com/releases/2019/01/190103152906.htm` (visited on 02/01/2020).

[8] Adam Alter. *Irresistible : the rise of addictive technology and the business of keeping us hooked*. The Penguin Group, 2017, p. 354. ISBN: 9781594206641. URL: `http://adamalterauthor.com/irresistible`.

[9] Julia Angwin et al. *Machine Bias*. 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing` (visited on 02/01/2020).

[10] Irene Solaiman et al. "Release Strategies and the Social Impacts of Language Models". In: *arXiv* (Aug. 2019). arXiv: `1908.09203`. URL: `http://arxiv.org/abs/1908.09203`.

[11] Mary L. Gray and Siddharth Suri. *Ghost work : how to stop Silicon Valley from building a new global underclass*. Eamon Dolan / Houghton Mifflin Harcourt, 2019, p. 254. ISBN: 1328566242. URL: `https://www.hmhco.com/shop/books/Ghost-Work/9781328566249`.

[12] Carl Benedikt Frey and Michael A Osborne. *The Future of Employment: How Susceptible Are Jobs to Computerisation?* Tech. rep. The Oxfort Marting Programme on Technology and Employment, 2013, p. 72. URL: `https://www.oxfordmartin.ox.ac.uk/downloads/academic/The%7B%5C_%7DFuture%7B%5C_%7Dof%7B%5C_%7DEmployment.pdf`.

[13]   Jess Whittlestone et al. "The role and limits of principles in AI ethics: Towards a focus on ten-
       sions". In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.
       Association for Computing Machinery, Inc, Jan. 2019, pp. 195–200. ISBN: 9781450363242. DOI:
       10.1145/3306618.3314289. URL: https://www.researchgate.net/publi
       cation/334378492%7B%5C_%7DThe%7B%5C_%7DRole%7B%5C_%7Dand%7B%5C
       %7DLimits%7B%5C%7Dof%7B%5C_%7DPrinciples%7B%5C_%7Din%7B%5C_%
       7DAI%7B%5C_%7DEthics%7B%5C_%7DTowards%7B%5C_%7Da%7B%5C_%7DFocu
       s%7B%5C_%7Don%7B%5C_%7DTensions.

[14]   Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In:
       Nature Machine Intelligence 1.9 (Sept. 2019), pp. 389–399. DOI: 10.1038/s42256-019-0088-2.
       URL: https://www.nature.com/articles/s42256-019-0088-2.

[15]   Beijing AI Principles Blog. *Beijing AI Principles*. 2019. URL: https://www.baai.ac.cn/
       blog/beijing-ai-principles (visited on 02/01/2020).

[16]   Yi Zeng, Enmeng Lu, and Cunqing Huangfu. "Linking Artificial Intelligence Principles". In: *arXiv*
       (Dec. 2018). arXiv: 1812.04814. URL: http://arxiv.org/abs/1812.04814.

[17]   Edelman. *2019 Edelman Trust Barometer | Edelman*. 2019. URL: https://www.edelman.
       com/research/2019-edelman-trust-barometer (visited on 02/01/2020).

[18]   Baobao Zhang and Allan Dafoe. *Artificial Intelligence: American Attitudes and Trends*. Tech. rep.
       Future of Humanity Institute, 2019. URL: https://governanceai.github.io/US-
       Public-Opinion-Report-Jan-2019/.

[19]   Gregory Travis. *How the Boeing 737 Max Disaster Looks to a Software Developer*. 2019. URL: http
       s://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-m
       ax-disaster-looks-to-a-software-developer (visited on 02/08/2020).

[20]   Jenny et al. Gesley. "Regulation of Artificial Intelligence in Selected Jurisdictions". In: (Jan.
       2019). URL: https://www.loc.gov/law/help/artificial-intelligence/
       regulation-artificial-intelligence.pdf.

[21]   E. L. Trist. *The evolution of socio-technical systems : a conceptual framework and an action research
       program*. Ontario Quality of Working Life Centre, 1981, p. 67. ISBN: 9780774362863. URL: http
       s://www.semanticscholar.org/paper/The-Evolution-of-Socio-Tech
       nical-Systems%7B%5C%7D3A-A-and-an-Trist/0ca460a31c%2093e5d7d
       1789be14e3cc30338f4ad4c.

[22]   Sundar Pichai. *AI at Google: our principles*. 2018. URL: https://www.blog.google/te
       chnology/ai/ai-principles/ (visited on 02/01/2020).

[23]   OpenAI. *OpenAI Charter*. 2018. URL: https://openai.com/charter/ (visited on 02/01/2020).

[24]   Microsoft. *AI Principles & Approach from Microsoft*. URL: https://www.microsoft.com
       /en-us/ai/our-approach-to-ai (visited on 02/01/2020).

[25]   Government of Canada. *Algorithmic Impact Assessment (AIA) - Canada.ca*. 2019. URL: https:/
       /www.canada.ca/en/government/system/digital-government/modern-
       emerging-technologies/responsible-use-ai/algorithmic-impact-a
       ssessment.html.

[26]   Treasury Board of Canada Secretariat. *Directive on Automated Decision-Making*. Tech. rep. Trea-
       sury Board of Canada Secreteriat, 2019. URL: https://www.tbs-sct.gc.ca/pol/d
       oc-eng.aspx?id=32592.

[27] Inioluwa Deborah Raji and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles". In: *arXiv* (Dec. 2019). arXiv: 1912.06166. URL: http://arxiv.org/abs/1912.06166.

[28] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *FAT '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. Oct. 2019. DOI: 10.1145/3287560.3287596. arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993%20http://dx.doi.org/10.1145/3287560.3287596.

[29] Timnit Gebru et al. "Datasheets for Datasets". In: *arXiv* (Mar. 2018). arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010.

[30] Cullen O'Keefe et al. "The Windfall Clause: Distributing the Benefits of AI for the Common Good". In: *arXiv* (Dec. 2019). arXiv: 1912.11595. URL: http://arxiv.org/abs/1912.11595.

[31] James M. Buchanan. "An Economic Theory of Clubs". In: *Economica* 32.125 (Feb. 1965), p. 1. ISSN: 00130427. DOI: 10.2307/2552442. URL: https://www.jstor.org/stable/2552442?seq=1.

[32] ETSI. *Intellectual Property Rights*. URL: https://www.etsi.org/intellectual-property-rights (visited on 02/01/2020).

[33] International Trade Forum. "Voluntary standards in developing Countries: The potential of voluntary standards and their role in international trade". In: (Mar. 2010). URL: http://www.tradeforum.org/Voluntary-Standards-in-Developing-Countries-The-Potential-of-Voluntary-Standards-and-their-Role-in-Inter%20national-Trade/.

[34] Future of Life Institute. *AI Principles*. 2017. URL: https://futureoflife.org/ai-principles/?cn-reloaded=1 (visited on 02/01/2020).

[35] Andy Pasztor and Alison Sider. *Internal Boeing Documents Show Cavalier Attitude to Safety*. 2020. URL: https://www.wsj.com/articles/internal-boeing-documents-show-cavalier-attitude-to-safety-11578627206 (visited on 02/06/2020).

[36] Inioluwa Deborah Raji et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing". In: *arXiv* (Jan. 2020). arXiv: 2001.00973. URL: http://arxiv.org/abs/2001.00973.

[37] David Engstrom and Daniel Ho. "Algorithmic Accountability in the Administrative State". In: *CSAS Working Papers* (Nov. 2019). URL: https://administrativestate.gmu.edu/wp-content/uploads/sites/29/2019/11/Engstrom-Ho-Algorithmic-Accountability-in-the-Administrative-State.pdf.

[38] Jan Barton. "Who Cares about Auditor Reputation?" In: *Contemporary Accounting Research* 22.3 (Oct. 2005), pp. 549–586. ISSN: 0823-9150. DOI: 10.1506/C27U-23K8-E1VL-20R0. URL: http://doi.wiley.com/10.1506/C27U-23K8-E1VL-20R0.

[39] Ryan Carrier. "Governance and Oversight Coming to AI and Automation: Independent Audit of AI Systems". In: *ACM blog* (Feb. 2019). URL: https://cacm.acm.org/blogs/blog-cacm/234723-governance-and-oversight-coming-to-ai-and-automation-independent-audit-of-ai-systems/fulltext.

[40] Jesse Hempel. "Want to Prove Your Business is Fair? Audit Your Algorithm". In: (May 2018). URL: https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm.

[41] Oren Etzioni and Michael Li. "Want to Prove Your Business is Fair? Audit Your Algorithm". In: (July 2019). URL: https://www.wired.com/story/ai-needs-to-be-audited.

[42] Jack Clark and Gillian K. Hadfield. "Regulatory Markets for AI Safety". In: *arXiv* (Dec. 2019). arXiv: 2001.00078. URL: http://arxiv.org/abs/2001.00078.

[43] Andrew Marshall et al. "Threat Modeling AI/ML Systems and Dependencies". In: (Nov. 2019). URL: https://docs.microsoft.com/en-us/security/threat-modeling-aiml.

[44] Ariel Herbert-Voss. *Dont Red Team AI Like a Chump*. 2019. URL: https://www.youtube.com/watch?v=ivlc1yPw76A (visited on 02/01/2020).

[45] Hackerone. *The Hacker-Powered Security Report 2017*. Tech. rep. Hackerone, 2017. URL: https://www.hackerone.com/sites/default/files/2017-06/The%20Hacker-Powered%20Security%20Report.pdf.

[46] J. B. Rubinovitz. *Bias Bounty Programs as a Method of Combatting Bias in AI*. 2018. URL: https://rubinovitz.com/post/187669528132/bias-bounty-programs-as-a-method-of-combatting (visited on 02/01/2020).

[47] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. ISSN: 0036-8075. URL: https://science.sciencemag.org/content/366/6464/447.

[48] Taylor Telford. *Apple Card algorithm sparks gender bias allegations against Goldman Sachs*. Nov. 2019. URL: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/ (visited on 02/06/2020).

[49] Dario Amodei et al. "Concrete Problems in AI Safety". In: *arXiv* (June 2016). arXiv: 1606.06565. URL: http://arxiv.org/abs/1606.06565.

[50] Warwick Ashford. *Bug bounties not a silver bullet, Katie Moussouris warns*. 2018. URL: https://www.computerweekly.com/news/252450337/Bug-bounties-not-a-silver-bullet-Katie-Moussouris-warns (visited on 02/08/2020).

[51] David Dao. *Awful AI*. URL: https://github.com/daviddao/awful-ai (visited on 02/01/2020).

[52] Victoria Krakovna. "Specification gaming examples in AI". In: (Apr. 2018). URL: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

[53] Peter Lee. "Learning from Tay's introduction". In: (Mar. 2016). URL: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

[54] Thomas Dietterich. "Robust Artificial Intelligence and Robust Human Organizations". In: *arXiv* (Mar. 2018). URL: https://arxiv.org/abs/1811.10840.

[55] Aviv Ovadya and Jess Whittlestone. "Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning". In: *arXiv* (July 2019). arXiv: 1907.11274. URL: http://arxiv.org/abs/1907.11274.

[56] Jan Leike et al. "AI Safety Gridworlds". In: *arXiv* (Nov. 2017). arXiv: 1711.09883. URL: http://arxiv.org/abs/1711.09883.

[57] Christian Szegedy et al. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2014. arXiv: 1312.6199. URL: https://arxiv.org/abs/1312.6199.

[58] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2015. arXiv: 1412.6572. URL: https://arxiv.org/abs/1412.6572.

[59] Alice Xiang and Inioluwa Deborah Raji. "On the Legal Compatibility of Fairness Definitions". In: *arXiv* (Nov. 2019). arXiv: 1912.00761. URL: http://arxiv.org/abs/1912.00761.

[60] Reuben Binns. "Fairness in Machine Learning: Lessons from Political Philosophy". In: *arXiv* (Dec. 2017). arXiv: 1712.03586. URL: http://arxiv.org/abs/1712.03586.

[61] Mark Koren, Anthony Corso, and Mykel Kochenderfer. "The Adaptive Stress Testing Formulation". In: *RSS 2019: Workshop on Safe Autonomy*. Freiburg, 2019. URL: https://openreview.net/pdf?id=rJgoNK-oaE.

[62] Ritchie Lee et al. "Adaptive stress testing of airborne collision avoidance systems". In: *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. Institute of Electrical and Electronics Engineers Inc., Oct. 2015. ISBN: 9781479989409. DOI: 10.1109/DASC.2015.7311613. URL: https://ieeexplore.ieee.org/document/7311613/versions.

[63] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* (June 2002). DOI: 10.1613/jair.953. arXiv: 1106.1813. URL: http://arxiv.org/abs/1106.1813%20http://dx.doi.org/10.1613/jair.953.

[64] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 7 (2016), pp. 1–5. arXiv: 1609.06570v1. URL: http://www.jmlr.org/papers/volume18/16-365/16-365.pdf.

[65] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *arXiv* (June 2017). arXiv: 1706.06083. URL: http://arxiv.org/abs/1706.06083.

[66] Matthias Hein and Maksym Andriushchenko. "Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation". In: *arXiv* (May 2017). arXiv: 1705.08475. URL: http://arxiv.org/abs/1705.08475.

[67] Mathias Lecuyer et al. "Certified Robustness to Adversarial Examples with Differential Privacy". In: *arXiv* (Feb. 2018). arXiv: 1802.03471. URL: http://arxiv.org/abs/1802.03471.

[68] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. "Certified Adversarial Robustness via Randomized Smoothing". In: *arXiv* (Feb. 2019). arXiv: 1902.02918. URL: http://arxiv.org/abs/1902.02918.

[69] Guy Katz et al. "Reluplex: An efficient smt solver for verifying deep neural networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10426 LNCS. Springer Verlag, 2017, pp. 97–117. ISBN: 9783319633862. DOI: 10.1007/978-3-319-63387-9_5. arXiv: 1702.01135. URL: https://arxiv.org/abs/1702.01135.

[70] Kevin Bonsor and Nathan Chandler. *How Black Boxes Work*. URL: `https://science.h owstuffworks.com/transport/flight/modern/black-box.htm` (visited on 02/02/2020).

[71] Mustafa Suleyman and Ben Laurie. *Trust, confidence and Verifiable Data Audit*. 2017. URL: `https ://deepmind.com/blog/article/trust-confidence-verifiable-data -audit` (visited on 02/02/2020).

[72] Joanna Bryson. "AI & Global Governance: No One Should Trust AI". In: (Nov. 2018). URL: `http s://cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.h tml`.

[73] Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *arXiv* (Nov. 2018). arXiv: `1811.10154`. URL: `htt p://arxiv.org/abs/1811.10154`.

[74] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *arXiv* (June 2016). arXiv: `1606.03490`. URL: `http://arxiv.org/abs/1606.03490`.

[75] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *arXiv* (Mar. 2017). arXiv: `1703.01365`. URL: `http://arxiv.org/abs/1703.01 365`.

[76] Mark Sendak et al. ""The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning". In: *arXiv* (Nov. 2019). arXiv: `1911.08089`. URL: `http://arxiv.org/ abs/1911.08089`.

[77] Berk Ustun, Alexander Spangher, and Yang Liu. "Actionable Recourse in Linear Classification". In: *FAT '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. Sept. 2019. DOI: `10.1145/3287560.3287566`. arXiv: `1809.06514`. URL: `http://arxiv .org/abs/1809.06514%20http://dx.doi.org/10.1145/3287560.328756 6`.

[78] Forough Poursabzi-Sangdeh et al. "Manipulating and Measuring Model Interpretability". In: *arXiv* (Feb. 2018). arXiv: `1802.07810`. URL: `http://arxiv.org/abs/1802.0781 0`.

[79] Neil Vigdor. *Apple Card Investigated After Gender Discrimination Complaints - The New York Times*. 2019. URL: `https://www.nytimes.com/2019/11/10/business/Apple-cred it-card-investigation.html` (visited on 02/02/2020).

[80] David Bau et al. "Semantic photo manipulation with a generative image prior". In: *ACM Transactions on Graphics* 38.4 (2019). ISSN: 15577368. DOI: `10.1145/3306346.3323023`. URL: `h ttp://ganpaint.io/Bau%7B%5C_%7Det%7B%5C_%7Dal%7B%5C_%7DSemanti c%7B%5C_%7DPhoto%7B%5C_%7DManipulation%7B%5C_%7Dpre%20print.pd f`.

[81] Fred Hohman et al. "Gamut: A design probe to understand how data scientists understand machine learning models". In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, May 2019. ISBN: 9781450359702. URL: `https://www.m icrosoft.com/en-us/research/uploads/prod/2019/01/19%7B%5C_%7Dg amut%7B%5C_%7Dchi.pdf`.

[82] Isaac Lage et al. "An Evaluation of the Human-Interpretability of Explanation". In: *arXiv* (Jan. 2019). arXiv: `1902.00006`. URL: `http://arxiv.org/abs/1902.00006`.

[83] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err". In: *Journal of Experimental Psychology: General* 144.1 (2015), pp. 114–126. ISSN: 00963445. DOI: `10.1037/xge0000033`. URL: `https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392%7B%5C&%7Dcontext=fnce%7B%5C_%7Dpapers`.

[84] Umang Bhatt et al. "Explainable Machine Learning in Deployment". In: *arXiv* (Sept. 2019). URL: `https://arxiv.org/abs/1909.06342`.

[85] Julius Adebayo et al. "Sanity Checks for Saliency Maps". In: *arXiv* (Oct. 2018). arXiv: `1810.03292`. URL: `http://arxiv.org/abs/1810.03292`.

[86] Aravindh Mahendran and Andrea Vedaldi. "Salient deconvolutional networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9910 LNCS. Springer Verlag, 2016, pp. 120–135. ISBN: 9783319464657. DOI: `10.1007/978-3-319-46466-4_8`. URL: `https://www.robots.ox.ac.uk/%7B~%7Dvedaldi/assets/pubs/mahendran16salient.pdf`.

[87] Pieter-Jan Kindermans et al. "The (Un)reliability of saliency methods". In: *arXiv* (Nov. 2017). arXiv: `1711.00867`. URL: `http://arxiv.org/abs/1711.00867`.

[88] Sara Hooker et al. "A Benchmark for Interpretability Methods in Deep Neural Networks". In: *arXiv* (June 2018). arXiv: `1806.10758`. URL: `http://arxiv.org/abs/1806.10758`.

[89] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain". In: *arXiv* (Aug. 2017). arXiv: `1708.06733`. URL: `http://arxiv.org/abs/1708.06733`.

[90] Juyeon Heo, Sunghwan Joo, and Taesup Moon. "Fooling Neural Network Interpretations via Adversarial Model Manipulation". In: *arXiv* (Feb. 2019). arXiv: `1902.02041`. URL: `http://arxiv.org/abs/1902.02041`.

[91] Dylan Slack et al. "How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods". In: *arXiv* (Nov. 2019). arXiv: `1911.02508`. URL: `http://arxiv.org/abs/1911.02508`.

[92] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-Augu. Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: `10.1145/2939672.2939778`. arXiv: `1602.04938`. URL: `https://arxiv.org/abs/1602.04938`.

[93] B Dimanov et al. "You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods". In: *European Conference on Artificial Intelligence*. 2020. URL: `http://ecai2020.eu/accepted-papers/`.

[94] Kamalika Chaudhuri and Claire Monteleoni. "Privacy-preserving logistic regression". In: *Neural Information Processing Systems Conference*. 2018. URL: `https://papers.nips.cc/paper/3486-privacy-preserving-logistic-regression`.

[95] Peter Kairouz et al. "Advances and Open Problems in Federated Learning". In: *arXiv* (Dec. 2019). arXiv: `1912.04977`. URL: `http://arxiv.org/abs/1912.04977`.

[96] Nicholas Carlini et al. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks". In: *arXiv* (Feb. 2018). arXiv: `1802.08232`. URL: `http://arxiv.org/abs/1802.08232`.

[97]   H. Brendan McMahan et al. "Learning Differentially Private Recurrent Language Models". In: *arXiv* (Oct. 2017). arXiv: `1710.06963`. URL: `http://arxiv.org/abs/1710.0696 3`.

[98]   Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 3876 LNCS. 2006, pp. 265–284. ISBN: 3540327312. URL: `https://p eople.csail.mit.edu/asmith/PS/sensitiv%20ity-tcc-final.pdf`.

[99]   Aaron Roth and Michael Kearns. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

[100] Rachel Cummings and Deven Desai. "The Role of Differential Privacy in GDPR Compliance". In: *FAT '18: Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2018. URL: `h ttps://piret.gitlab.io/fatrec2018/program/fatrec2018-cummings. pdf`.

[101] Mihailo Isakov et al. "Survey of Attacks and Defenses on Edge-Deployed Neural Networks". In: *arXiv* (Nov. 2019). arXiv: `1911.11932`. URL: `http://arxiv.org/abs/1911.1193 2`.

[102] Nathan Dowlin et al. "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy". In: *Proceedings of the 33rd International Conference on Machine Learning*. 2016. URL: `http://sealcrypto.codeplex.com`.

[103] Tim Hwang. "Computational Power and the Social Impact of Artificial Intelligence". In: *SSRN Electronic Journal* (Apr. 2018). DOI: `10.2139/ssrn.3147971`. arXiv: `1803.08971`. URL: `https://arxiv.org/abs/1803.08971`.

[104] Rich Sutton. *The Bitter Lesson*. 2019. URL: `http://www.incompleteideas.net/Inc Ideas/BitterLesson.html` (visited on 02/02/2020).

[105] Rodney Brooks. *A Better Lesson*. 2019. URL: `https://rodneybrooks.com/a-better -lesson/` (visited on 02/02/2020).

[106] Huili Chen et al. "DeepAttest: An end-to-end attestation framework for deep neural networks". In: *Proceedings of the International Symposium on Computer Architecture*. Institute of Electrical and Electronics Engineers Inc., June 2019, pp. 487–498. ISBN: 9781450366694. URL: `https:/ /www.microsoft.com/en-us/research/uploads/prod/2019/05/DeepAtt est.pdf`.

[107] Zahra Ghodsi, Tianyu Gu, and Siddharth Garg. "SafetyNets: Verifiable Execution of Deep Neural Networks on an Untrusted Cloud". In: *Conference on Neural Information Processing Systems*. 2017. URL: `https://papers.nips.cc/paper/7053-safetynets-verifiabl e-execution-of-deep-neural-networks-on-an-untrusted-cloud.pdf`.

[108] Seyyed Ahmad Javadi et al. "Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'". In: *arXiv* (Jan. 2020). arXiv: `2001.09723`. URL: `http://arxiv.org/abs/200 1.09723`.

[109] Frank Mckeen et al. "Innovative Instructions and Software Model for Isolated Execution". In: *HASP '13: Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy*. 2013. DOI: `10.1145/2487726.2488368`. URL: `https://www. eit.lth.se/fileadmin/eit/courses/eitn50/Literature/hasp-2013- innovative-instructions-and-software-model-for-isolated-execu tion.pdf`.

[110] Jo Van Bulck et al. "Foreshadow: Extracting the Keys to the Intel {SGX} Kingdom with Transient Out-of-Order Execution". In: *27th {USENIX} Security Symposium ({USENIX} Security 18)*. Baltimore, MD: {USENIX} Association, Aug. 2018, 991{\textendash}1008. ISBN: 978-1-939133-04-5. URL: `https://www.usenix.org/conference/usenixsecurity18/presentation/bulck`.

[111] Mark Seaborn and Thomas Dullien. *Exploiting the DRAM rowhammer bug to gain kernel privileges How to cause and exploit single bit errors*. 2015. URL: `https://www.blackhat.com/docs/us-15/materials/us-15-Seaborn-Exploiting-The-DRAM-Rowhammer-Bug-To-Gain-Kernel-Privileges.pdf` (visited on 02/02/2020).

[112] Nick Hynes, Raymond Cheng, and Dawn Song. "Efficient Deep Learning on Multi-Source Private Data". In: *arXiv* (July 2018). arXiv: `1807.06689`. URL: `http://arxiv.org/abs/1807.06689`.

[113] Christian Priebe et al. "SGX-LKL: Securing the Host OS Interface for Trusted Execution". In: *arXiv* (Aug. 2019). arXiv: `1908.11143`. URL: `http://arxiv.org/abs/1908.11143`.

[114] Ehsan Hesamifard et al. "Privacy-preserving Machine Learning as a Service". In: *Proceedings on Privacy Enhancing Technologies*. 2018, pp. 123–142. arXiv: `1803.05961`. URL: `https://petsymposium.org/2018/files/papers/issue3/popets-2018-0024.pdf`.

[115] Stavros Volos, Kapil Vaswani, and Rodrigo Bruno. "Graviton: Trusted Execution Environments on GPUs". In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. OSDI'18. USA: USENIX Association, 2018, pp. 681–696. ISBN: 9781931971478. URL: `https://www.microsoft.com/en-us/research/uploads/prod/2018/09/Graviton-Trusted-Execution-Environments-on-GPUs.pdf`.

[116] Florian Tramèr and Dan Boneh. "Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware". In: *arXiv* (June 2018). arXiv: `1806.03287`. URL: `http://arxiv.org/abs/1806.03287`.

[117] Zhongshu Gu et al. "YerbaBuena: Securing Deep Learning Inference Data via Enclave-based Ternary Model Partitioning". In: *arXiv* (July 2018). arXiv: `1807.00969`. URL: `http://arxiv.org/abs/1807.00969`.

[118] Dario Amodei et al. *AI and Compute*. 2018. URL: `https://openai.com/blog/ai-and-compute/` (visited on 02/02/2020).

[119] Lily Hay Newman. *Google Is Helping Design an Open Source, Ultra-Secure Chip | WIRED*. 2019. URL: `https://www.wired.com/story/open-titan-open-source-secure-enclave/` (visited on 02/02/2020).

[120] Cisco. *Cisco Global Cloud Index: Forecast and Methodology, 2016-2021 White Paper*. Tech. rep. Cisco, 2018. URL: `https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html`.

[121] Roy Schwartz et al. "Green AI". In: *arXiv* (July 2019). arXiv: `1907.10597`. URL: `http://arxiv.org/abs/1907.10597`.

[122] Christopher Berner et al. "Dota 2 with Large Scale Deep Reinforcement Learning". In: *arXiv* (Dec. 2019). arXiv: `1912.06680`. URL: `http://arxiv.org/abs/1912.06680`.

[123] J Vitek and T Kalibera. "Repeatability, reproducibility and rigor in systems research". In: *2011 Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT)*. Oct. 2011, pp. 33–38. DOI: `10.1145/2038642.2038650`. URL: `https://ieeexplore.ieee.org/document/6064509`.

[124] T Hoefler and R Belli. "Scientific benchmarking of parallel computing systems: twelve ways to tell the masses when reporting performance results". In: *SC '15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. Nov. 2015, pp. 1–12. DOI: `10.1145/2807591.2807644`. URL: `https://htor.inf.ethz.ch/publications/img/hoefler-scientific-benchmarking.pdf`.

[125] Steve Lohr. *At Tech's Leading Edge, Worry About a Concentration of Power*. 2019. URL: `https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html` (visited on 02/02/2020).

[126] Martijn Rasser et al. *The American AI Century: A Blueprint for Action*. Tech. rep. Center for a New American Security, Dec. 2019. URL: `https://www.cnas.org/publications/reports/the-american-ai-century-a-blueprint-for-action`.

[127] John Etchemendy and Fei-Fei Li. "National Research Cloud: Ensuring the Continuation of American Innovation". In: (Mar. 2020). URL: `https://hai.stanford.edu/news/national-research-cloud-ensuring-continuation-american-innovation`.

[128] Susanne Barth and Menno D.T. de Jong. *The privacy paradox - Investigating discrepancies between expressed privacy concerns and actual online behavior - A systematic literature review*. Nov. 2017. DOI: `10.1016/j.tele.2017.04.013`. URL: `https://www.sciencedirect.com/science/article/pii/S0736585317302022`.

[129] Ehsan Toreini et al. "The relationship between trust in AI and trustworthy machine learning technologies". In: *arXiv* (Nov. 2019). arXiv: `1912.00782`. URL: `http://arxiv.org/abs/1912.00782`.

[130] Mike Ananny and Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". In: *New Media & Society* 20.3 (Mar. 2018), pp. 973–989. ISSN: 1461-4448. DOI: `10.1177/1461444816676645`. URL: `http://journals.sagepub.com/doi/10.1177/1461444816676645`.

[131] Joshua A. Kroll et al. "Accountable Algorithms". In: *University of Pennsylvania Law Review* 165 (2017). URL: `https://papers.ssrn.com/sol3/papers.cfm?abstract%7B%5C_%7Did=2765268%7B%5C#%7D%7B%5C#%7D`.

[132] Partnership on AI. "Human-AI Collaboration Trust Literature Review - Key Insights and Bibliography". In: (Sept. 2019). URL: `https://www.partnershiponai.org/human-ai-collaboration-trust-literature-review-key-insights-and-bibliography/`.

[133] High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. Tech. rep. High-Level Expert Group on AI, 2019. URL: `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

[134] Google PAIR. *Explainability + Trust*. URL: `https://pair.withgoogle.com/chapter/explainability-trust/` (visited on 02/05/2020).

[135] Matthew Arnold et al. "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity". In: *arXiv* (Aug. 2018). arXiv: `1808.07261`. URL: `http://arxiv.org/abs/1808.07261`.

[136] Roger C Mayer, James H Davis, and F David Schoorman. "An Integrative Model of Organizational Trust". In: *The Academy of Management Review* 20.3 (1995), pp. 709–734. ISSN: 03637425. URL: http://www.jstor.org/stable/258792.

[137] D Gambetta. *Trust: Making and Breaking Cooperative Relations*. B. Blackwell, 1988. ISBN: 978063 1155065. URL: https://books.google.com.tr/books?id=97VmQgAACAAJ.

[138] ANDREW J COE and JANE VAYNMAN. "Why Arms Control Is So Rare". In: *American Political Science Review* (2019), pp. 1–14. DOI: 10.1017/S000305541900073X. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/why-arms-control-is-so-rare/BAC79354627F72CDDD%20B102FE82889B8A.

[139] Suzanne Massie and Nicholas (ed.) Daniloff. "Suzanne Massie, Trust but Verify: Reagan, Russia, and Me. Rockland, ME: Maine Authors Publishing, 2013. 380 pp." In: *Journal of Cold War Studies* 18.4 (2016), pp. 225–228. DOI: 10.1162/JCWS_r_00693. URL: https://doi.org/10.1162/JCWS%7B%5C_%7Dr%7B%5C_%7D00693.

[140] Robin Bloomfield et al. "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy". In: *Computer* 52.9 (Sept. 2019), pp. 82–89. ISSN: 15580814. DOI: 10.1109/MC.2019.2914775. URL: https://openaccess.city.ac.uk/id/eprint/23295/1/bloomfield%7B%5C_%7DpreIEEE%7B%5C_%7Dassuring%7B%5C_%7Dautonomy%7B%5C_%7Dv01i.pdf.

[141] Office for Nuclear Regulation (ONR). *ONR Guide*. Tech. rep. Office for Nuclear Regulation, 2019. URL: http://www.onr.org.uk/operational/tech%7B%5C_%7Dasst%7B%5C_%7Dguides/ns-tast-gd-051.pdf.

[142] Uber. *Uber ATG Safety Case*. 2019. URL: https://uberatg.com/safetycase/ (visited on 02/01/2020).

[143] Xingyu Zhao et al. "A Safety Framework for Critical Systems Utilising Deep Neural Networks". In: *arXiv* (Mar. 2020). URL: https://arxiv.org/abs/2003.05311.

[144] Stephen Edelston Toulmin, joint author Rieke Richard D., and joint author Janik Allan. *An introduction to reasoning*. New York Macmillan, 1979. ISBN: 0024210307. URL: http://openlibrary.org/books/OL4723061M.

[145] Donald G. (ed) Brennan. *Arms Control, Disarmament, and National Security*. G. Braziller, 1961. URL: https://www.jstor.org/stable/pdf/20025353.pdf?seq=1.

[146] Bernard Brodie. "HEDLEY BULL. The Control of the Arms Race: Disarmament and Arms Control in the Missile Age. (Studies in Inter national Security, II.) Pp. 215. New York: Frederick A. Praeger for the Institute for Strategic Studies, 1961. $3.95". In: *The ANNALS of the American Academy of Political and Social Science* 341.1 (1962), pp. 115–116. DOI: 10.1177/000271626234100116. URL: https://doi.org/10.1177/000271626234100116.

[147] Richard Dean Burns. *Encyclopedia of arms control and disarmament*. Encyclopedia of Arms Control and Disarmament v. 1-3. Scribner's, 1993. ISBN: 9780684192819. URL: https://books.google.com.tr/books?id=f42yAAAAIAAJ.

[148] Richard Dean Burns. *The Evolution of Arms Control: From Antiquity to the Nuclear Age*. Rowman & Littlefield, 2013, p. 264. ISBN: 1442223790. URL: https://rowman.com/ISBN/9781442223790/The-Evolution-of-Arms-Control-From-Antiquity-to-the-Nuclear-Age.

[149] Jozef Goldblat. *Agreements for Arms Control: A Critical Survey*. Taylor & Francis, 1982, p. 387. ISBN: 0-85066-229-X. URL: `https://www.sipri.org/publications/1982/agreements-arms-control-critical-survey`.

[150] Tim Caughley. *Nuclear Disarmament Verification: Survey of Verification Mechanisms*. Tech. rep. United Nations Institute for Disarmament Research, 2016, p. 48. URL: `https://unidir.org/publication/nuclear-disarmament-verification-survey-verif%20ication-mechanisms`.

[151] Nancy W. Gallagher. *The politics of verification*. Johns Hopkins University Press, 2003, p. 311. ISBN: 9780801877391. URL: `https://www.tandfonline.com/doi/abs/10.1080/13523269708404165`.

[152] John H. Herz. "Political Realism and Political Idealism: A Study in Theories and Realities". In: *American Political Science Review* 46.2 (June 1952), pp. 550–552. ISSN: 0003-0554. DOI: `10.2307/1950848`. URL: `https://www.cambridge.org/core/journals/american-political-science-review/article/political-realism-and-political-idealism-a-study-in-theories-and-realities-by-john-h-herz-chicago-university-of-chicago-press-1951-pp-xii-275-375/8DA0D7344AD6FC86873EA5A4935`.

[153] Robert Jervis. "Cooperation Under the Security Dilemma". In: *World Politics* 30.2 (1978), pp. 167–214. ISSN: 00438871, 10863338. URL: `http://www.jstor.org/stable/2009958`.

[154] Robert Jervis. "Arms Control, Stability, and Causes of War". In: *Political Science Quarterly* 108.2 (1993), pp. 239–253. ISSN: 00323195. URL: `http://www.jstor.org/stable/2152010`.

[155] Organisation for the Prohibition of Chemical Weapons. *Chemical Weapons Convention*. Tech. rep. Organisation for the Prohibition of Chemical Weapons, 1993. URL: `https://www.opcw.org/chemical-weapons-convention`.

[156] Frederic Joseph Brown. *Chemical Warfare: A Study in Restraints*. Princeton University Press, 1968, p. 376. ISBN: 9780691069012. URL: `http://www.jstor.org/stable/j.ctt183ph9d`.

[157] Carol Barner-Barry. "The Diplomacy of Biological Disarmament: Vicissitudes of a Treaty in Force". In: *Politics and the Life Sciences* 9.1 (1990), pp. 149–149. DOI: `10.1017/S0730938400010352`. URL: `https://www.cambridge.org/core/journals/politics-and-the-life-sciences/article/diplomacy-of-biological-disarmament-vicissitudes-of-a-treaty-in-force-nicholas-a-sims-new-york-st-martins-press1988/58994CAE4E3EF4EEF484F5313FCAD83E`.

[158] Kenneth R Rutherford. "The Evolving Arms Control Agenda: Implications of the Role of NGOs in Banning Antipersonnel Landmines". In: *World Politics* 53.1 (2000), pp. 74–114. ISSN: 00438871, 10863338. URL: `http://www.jstor.org/stable/25054137`.

[159] United Nations. *Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction*. Oslo, 1997. URL: `https://treaties.un.org/Pages/ViewDetails.aspx?src=IND%7B%5C&%7Dmtdsg%7B%5C_%7Dno=XXVI-5%7B%5C&%7Dchapter=26%7B%5C&%7Dclang=%7B%5C_%7Den`.

[160] United Nations. *The Convention on Cluster Munitions*. 2008. URL: `https://www.clusterc%20onvention.org/`.

[161] Heather Roff and Richard Moyes. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons". In: *Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons*. 2016. URL: http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf.

[162] Atomic Heritage Foundation. *Non-Proliferation, Limitation, and Reduction*. 2017. URL: https://www.atomicheritage.org/history/non-proliferation-limitation-and-reduction (visited on 02/02/2020).

[163] United States and USSR. *Treaty On The Limitation of Anti-Ballistic Missile Systems*. 1972. URL: https://www.nti.org/learn/treaties-and-regimes/treaty-limitat%20ion-anti-ballistic-missile-systems-abm-treaty/.

[164] Rebecca Crootof. "The Killer Robots Are Here: Legal and Policy Implications". In: *Cardozo law review* 36 (2015), p. 1837. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract%7B%5C_%7Did=2534567.

[165] Sean Watts. "Autonomous Weapons: Regulation Tolerant or Regulation Resistant?" In: *SSRN Electronic Journal* (Nov. 2015). DOI: 10.2139/ssrn.2681283. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract%7B%5C_%7Did=2681283.

[166] Paul Scharre. *Army of None*. W. W. Norton & Company, 2019, p. 448. URL: https://wwnorton.com/books/Army-of-None/.

[167] Kenneth Payne. "Artificial Intelligence: A Revolution in Strategic Affairs?" In: *Survival* 60.5 (2018), pp. 7–32. DOI: 10.1080/00396338.2018.1518374. URL: https://doi.org/10.1080/00396338.2018.1518374.

[168] Miles Brundage et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". In: *arXiv* (Feb. 2018). arXiv: 1802.07228. URL: http://arxiv.org/abs/1802.07228.

[169] International Committe of the Red Cross. *Report of the ICRC Expert Meeting on 'Autonomous weapon systems: technical, military, legal and humanitarian aspects'*. Tech. rep. The Red Cross, 2014. URL: https://www.icrc.org/en/doc/assets/files/2014/expert-meeting-autonomous-weapons-icrc-report-2014-05-09.pdf.

[170] Nehal Bhuta et al., eds. *Autonomous weapons systems : law, ethics, policy*. Cambridge University Press, 2016, p. 410. ISBN: 9781316607657. URL: https://www.cambridge.org/tr/academic/subjects/law/humanitarian-law/autonomous-weapons-systems-law-ethics-policy?format=PB.

[171] Greg Allen and Taniel Chan. *Artificial Intelligence and National Security*. Tech. rep. Belfer Center for Science and International Affairs, Harvard Kennedy School, 2017. URL: https://www.belfercenter.org/publication/artificial-intelligence-and-nati%20onal-security.

[172] Andrew Imbrie and Elsa B. Kania. *AI Safety, Security, and Stability Among Great Powers*. Tech. rep. Center for Security and Emerging Technology, 2019, p. 25. URL: https://cset.georgetown.edu/wp-content/uploads/AI-Safety-Security-and-Stability-Among-the-Great-Powers.pdf.

[173] Richard Moyes. "Key elements of meaningful human control". In: *Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons*. 2016. URL: http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf.

[174] Heather Roff. "Autonomous weapon systems: Evaluating the capacity for 'meaningful human control' in weapon review processes". In: *Convention on Certain Conventional Weapons Group of Governmental Experts meeting on Lethal Autonomous Weapons Systems*. 2017. URL: `http://www.article36.org/wp-content/uploads/2013/06/Evaluating-human-control-1.pdf`.

[175] Haydn Belfield. "Activism by the AI Community: Analysing Recent Achievements and Future Prospects". In: *arXiv* (Jan. 2020). arXiv: `2001.06528`. URL: `http://arxiv.org/abs/2001.06528`.

[176] Charli Carpenter. "*Lost*" *Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Cornell University Press, 2014. ISBN: 9780801448850. URL: `http://www.jstor.org/stable/10.7591/j.ctt5hh0r5`.

[177] Serif Onur Bahcecik. "Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames". In: *Global Policy* 10.3 (2019), pp. 365–369. DOI: `10.1111/1758-5899.12671`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12671`.

[178] Maaike Verbruggen. "The Role of Civilian Innovation in the Development of Lethal Autonomous Weapon Systems". In: *Global Policy* 10.3 (2019), pp. 338–342. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12663`.

[179] Kerstin Vignard. *Manifestos and open letters: Back to the future?* 2018. URL: `https://thebull%20etin.org/2018/04/manifestos-and-open-letters-back-to-the-future/` (visited on 02/03/2020).

[180] Kara Frederick. *The civilian private sector: part of a new arms control regime?* 2019. URL: `https://www.orfonline.org/expert-speak/the-civilian-private-sector-part-of-a-new-arms-control-regime-57345/` (visited on 02/03/2020).

[181] Future of Life Institute. *Open Letter on Autonomous Weapons*. Tech. rep. Future of Life Institute, 2015. URL: `https://futureoflife.org/open-letter-autonomous-weapons/`.

[182] Emanuel Adler. "The emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control". In: *International Organization* 46.1 (1992), pp. 101–145. DOI: `10.1017/S0020818300001466`. URL: `https://www.cambridge.org/core/journals/international-organization/article/emergence-of-cooperation-national-epistemic-communities-and-the-international-evolution-of-the-idea-of-nuclear-arms-control/AD5AB338380EC8691C621B351BC11CE3`.

[183] Peter M. Haas. *Introduction: Epistemic Communities and International Policy Coordination*. 1992. DOI: `10.2307/2706951`. URL: `https://www.jstor.org/stable/2706951`.

[184] *Standard Oil Co. of New Jersey v. United States :: 221 U.S. 1*. May 1911. URL: `https://supreme.justia.com/cases/federal/us/221/1/`.

[185] William E. Kovacic and Carl Shapiro. "Antitrust Policy: A Century of Economic and Legal Thinking". In: *Journal of Economic Perspectives* 14.1 (2000), pp. 43–60. URL: `https://perma.cc/N6EF-NGD3`.

[186] B. Y. Orbach. "THE ANTITRUST CONSUMER WELFARE PARADOX". In: *Journal of Competition Law and Economics* 7.1 (Mar. 2011), pp. 133–164. ISSN: 1744-6414. URL: `https://academic.oup.com/jcle/article-lookup/doi/10.1093/joclec/nhq019`.

[187] Lina M Khan. "Amazon's Antitrust Paradox". In: *The Yale Law Journal* 126 (2017), pp. 710–805. URL: `http://www.internetretailer.com/2016/ol/28/amazon-sales-climb-22`.

[188] Astead W. Herndon. *Elizabeth Warren Proposes Breaking Up Tech Giants Like Amazon and Facebook*. 2019. URL: `https://www.nytimes.com/2019/03/08/us/politics/elizabeth-warren-amazon.html` (visited on 02/03/2020).

[189] *Arizona v. Maricopa County Med. Soc'y :: 457 U.S. 332*. June 1982. URL: `https://supreme.justia.com/cases/federal/us/457/332/`.

[190] Practical Law Antitrust. *US Antitrust Laws: Overview*. URL: `https://uk.practicallaw.thomsonreuters.com/9-204-0472`.

[191] *State Oil Co. v. Khan :: 522 U.S. 3*. Nov. 1997. URL: `https://supreme.justia.com/cases/federal/us/522/3/`.

[192] *FTC v. Indiana Fed'n of Dentists :: 476 U.S. 447*. June 1986. URL: `https://supreme.justia.com/cases/federal/us/476/447/`.

[193] *National Soc'y of Prof. Engineers v. United States :: 435 U.S. 679*. Apr. 1978. URL: `https://supreme.justia.com/cases/federal/us/435/679/`.

[194] Federal Trade Commision and U.S. Department of Justice. *Antitrust Guidelines for Collaborations Among Competitors*. Tech. rep. Federal Trade Commision, Apr. 2000. URL: `https://www.ftc.gov/sites/default/files/documents/public%7B%5C_%7Devents/joint-venture-hearings-antitrust-guidelines-collaboration-among-competitors/ftcdojguidelines-2.pdf`.

[195] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. "Interpreting Blackbox Models via Model Extraction". In: *arXiv* (May 2017). arXiv: `1705.08504`. URL: `http://arxiv.org/abs/1705.08504`.

[196] Luca Pulina and Armando Tacchella. "An Abstraction-Refinement Approach to Verification of Artificial Neural Networks". In: *CEUR Workshop Proceedings*. 2010, pp. 243–257. URL: `http://link.springer.com/10.1007/978-3-642-14295-6%7B%5C_%7D24`.

[197] Changliu Liu et al. "Algorithms for Verifying Deep Neural Networks". In: *arXiv* (Mar. 2019). arXiv: `1903.06758`. URL: `http://arxiv.org/abs/1903.06758`.

[198] Daniel Selsam, Percy Liang, and David L. Dill. "Developing Bug-Free Machine Learning Systems With Formal Mathematics". In: *arXiv* (June 2017). arXiv: `1706.08605`. URL: `http://arxiv.org/abs/1706.08605`.

[199] C Hutchison et al. "Robustness Testing of Autonomy Software". In: *2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP)*. May 2018, pp. 276–285. URL: `https://users.ece.cmu.edu/%7B~%7Dkoopman/pubs/hutchison18%7B%5C_%7Dicse%7B%5C_%7Drobustness%7B%5C_%7Dtesting%7B%5C_%7Dautonomy%7B%5C_%7Dsoftware.pdf`.

[200] Joe Gibbs Politz et al. "Python: The Full Monty". In: *Proceedings of the 2013 ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages & Applications*. OOPSLA '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 217–232. ISBN: 9781450323741. DOI: `10.1145/2509136.2509536`. URL: `https://doi.org/10.1145/2509136.2509536`.

[201] Philippa Anne Gardner, Sergio Maffeis, and Gareth David Smith. "Towards a Program Logic for JavaScript". In: *SIGPLAN Not.* 47.1 (Jan. 2012), pp. 31–44. ISSN: 0362-1340. URL: `https://doi.org/10.1145/2103621.2103663`.

[202] Lindsey Kuper et al. "Toward Scalable Verification for Safety-Critical Deep Networks". In: *arXiv* (Jan. 2018). arXiv: `1801.05950`. URL: `http://arxiv.org/abs/1801.05950`.

[203] Pang Wei Koh and Percy Liang. "Understanding Black-box Predictions via Influence Functions". In: *arXiv* (Mar. 2017). arXiv: `1703.04730`. URL: `http://arxiv.org/abs/1703.04730`.

[204] Rajiv Khanna et al. "Interpreting Black Box Predictions using Fisher Kernels". In: *arXiv* (Oct. 2018). arXiv: `1810.10118`. URL: `http://arxiv.org/abs/1810.10118`.

[205] Boris Sharchilev et al. "Finding Influential Training Samples for Gradient Boosted Decision Trees". In: *arXiv* (Feb. 2018). arXiv: `1802.06640`. URL: `http://arxiv.org/abs/1802.06640`.

[206] Chih-Kuan Yeh et al. "Representer Point Selection for Explaining Deep Neural Networks". In: *Conference on Neural Information Processing Systems*. 2018. URL: `https://papers.nips.cc/paper/8141-representer-point-selection-for-explaining-deep-neural-networks.pdf`.

[207] Quanshi Zhang et al. "Interpreting CNNs via Decision Trees". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019. URL: `https://arxiv.org/abs/1802.00121`.

[208] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. "Interpretable decision sets: A joint framework for description and prediction". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-Augu. Association for Computing Machinery, Aug. 2016, pp. 1675–1684. ISBN: 9781450342322. URL: `https://dl.acm.org/doi/10.1145/2939672.2939874`.

[209] Sarah Tan et al. "Learning Global Additive Explanations for Neural Nets Using Model Distillation". In: *arXiv* (Jan. 2018). arXiv: `1801.08640`. URL: `http://arxiv.org/abs/1801.08640`.

[210] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph". In: *arXiv* (Aug. 2017). arXiv: `1708.01785`. URL: `http://arxiv.org/abs/1708.01785`.

[211] Wieland Brendel and Matthias Bethge. "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet". In: *arXiv* (Mar. 2019). arXiv: `1904.00760`. URL: `http://arxiv.org/abs/1904.00760`.

[212] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. "Interpretable Convolutional Neural Networks". In: *arXiv* (Oct. 2017). arXiv: `1710.00935`. URL: `http://arxiv.org/abs/1710.00935`.

[213] Xi Chen et al. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". In: *arXiv* (June 2016). arXiv: `1606.03657`. URL: `http://arxiv.org/abs/1606.03657`.

[214] Irina Higgins et al. "β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK". In: *International Conference on Learning Representations*. 2017. URL: `https://openreview.net/references/pdf?id=Sy2fzU9gl`.

[215] Abhinav Verma et al. "Programmatically Interpretable Reinforcement Learning". In: *arXiv* (Apr. 2018). arXiv: `1804.02477`. URL: `http://arxiv.org/abs/1804.02477`.

[216] Andrew Slavin Ross and Finale Doshi-Velez. "Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients". In: *arXiv* (Nov. 2017). arXiv: 1711.09404. URL: http://arxiv.org/abs/1711.09404.

[217] Saleema Amershi et al. "Modeltracker: Redesigning performance analysis tools for machine learning". In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2015-April. Association for Computing Machinery, Apr. 2015, pp. 337–346. ISBN: 9781450331456. DOI: 10.1145/2702123.2702509. URL: https://www.microsoft.com/en-us/research/publication/modeltracker-redesigning-performance-ana%20lysis-tools-for-machine-learning/.

[218] Tongshuang Wu et al. "Errudite: Scalable, Reproducible, and Testable Error Analysis". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 747–763. URL: https://www.aclweb.org/anthology/P19-1073.

[219] Jesse Vig. "A Multiscale Visualization of Attention in the Transformer Model". In: *arXiv* (June 2019). arXiv: 1906.05714. URL: http://arxiv.org/abs/1906.05714.

[220] David Bau et al. "GaN dissection: Visualizing and understanding generative adversarial networks". In: *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, 2019. arXiv: 1811.10597. URL: https://arxiv.org/abs/1811.10597.

[221] Shan Carter et al. "Activation Atlas". In: *Distill* 4.3 (Mar. 2019). ISSN: 2476-0757. URL: https://distill.pub/2019/activation-atlas.

[222] Jost Tobias Springenberg et al. "Striving for Simplicity: The All Convolutional Net". In: *arXiv* (Dec. 2014). arXiv: 1412.6806. URL: http://arxiv.org/abs/1412.6806.

[223] Matthew D. Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, 2014, pp. 818–833. ISBN: 9783319105895. DOI: 10.1007/978-3-319-10590-153. arXiv: 1311.2901. URL: https://cs.nyu.edu/%7B~%7Dfergus/papers/zeilerECCV2014.pdf.

[224] Ruth Fong and Andrea Vedaldi. "Interpretable Explanations of Black Boxes by Meaningful Perturbation". In: *IEEE International Conference on Computer Vision*. Apr. 2017. arXiv: 1704.03296. URL: http://arxiv.org/abs/1704.03296%20http://dx.doi.org/10.1109/ICCV.2017.371.

[225] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. "Understanding Deep Networks via Extremal Perturbations and Smooth Masks". In: *IEEE International Conference on Computer Vision*. 2019. URL: https://arxiv.org/abs/1910.08485.

[226] Piotr Dabkowski and Yarin Gal. "Real Time Image Saliency for Black Box Classifiers". In: *Advances in Neural Information Processing Systems 30*. Ed. by I Guyon et al. Curran Associates, Inc., 2017, pp. 6967–6976. URL: http://papers.nips.cc/paper/7272-real-time-image-saliency-for-black-box-classifiers.pdf.

[227] Vitali Petsiuk, Abir Das, and Kate Saenko. "RISE: Randomized Input Sampling for Explanation of Black-box Models". In: *arXiv* (June 2018). arXiv: 1806.07421. URL: http://arxiv.org/abs/1806.07421.

[228] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv* (Dec. 2013). arXiv: 1312.6034. URL: http://arxiv.org/abs/1312.6034.

[229] Himabindu Lakkaraju et al. "Interpretable & Explorable Approximations of Black Box Models". In: *Fairness, Accountability, and Transparency in Machine Learning*. 2017. arXiv: 1707.01154v1. URL: https://arxiv.org/abs/1707.01154.

[230] Anh Nguyen et al. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks". In: *Advances in Neural Information Processing Systems 29*. Ed. by D D Lee et al. Curran Associates, Inc., 2016, pp. 3387–3395. URL: http://papers.nips.cc/paper/6519-synthesizing-the-preferred-inputs-for-neurons-in-neural-networks-via-deep-generator-networks.pdf.

[231] David Bau et al. "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *arXiv* (Apr. 2017). arXiv: 1704.05796. URL: http://arxiv.org/abs/1704.05796.

[232] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization". In: *Distill* 2.11 (Nov. 2017). ISSN: 2476-0757. DOI: 10.23915/distill.00007. URL: https://distill.pub/2017/feature-visualization.

[233] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep Image Prior". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2018. URL: https://arxiv.org/abs/1711.10925.

[234] Chris Olah et al. "The Building Blocks of Interpretability". In: *Distill* 3.3 (Mar. 2018). ISSN: 2476-0757. DOI: 10.23915/distill.00010. URL: https://distill.pub/2018/building-blocks.

[235] Ari S. Morcos et al. "On the importance of single directions for generalization". In: *arXiv* (Mar. 2018). arXiv: 1803.06959. URL: http://arxiv.org/abs/1803.06959.

[236] Karel Lenc and Andrea Vedaldi. "Understanding image representations by measuring their equivariance and equivalence". In: *arXiv* (Nov. 2014). arXiv: 1411.5908. URL: http://arxiv.org/abs/1411.5908.

[237] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[238] Carrie J. Cai et al. "Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making". In: *arXiv* (Feb. 2019). arXiv: 1902.02960. URL: http://arxiv.org/abs/1902.02960.

Appendices

I Workshop and Report Writing Process

This report began as an effort to identify areas for productive work related to trust in AI development. The project began in earnest with an interdisciplinary expert workshop in San Francisco in April of 2019, which brought together participants from academia, industry labs, and civil society organizations.[82] As discussed below, during the writing process, we shifted our focus to verifiable claims in particular, rather than trust more broadly.

Workshop attendees are listed below in alphabetical order:

- Amanda Askell
- Andrew Critch
- Andrew Lohn
- Andrew Reddie
- Andrew Trask
- Ben Garfinkel
- Brian Tse
- Catherine Olsson
- Charina Chou
- Chris Olah
- David Luan
- Dawn Song
- Emily Oehlsen
- Eric Sigler
- Genevieve Fried
- Gillian Hadfield
- Heidy Khlaaf
- Helen Toner

- Ivan Vendrov
- Jack Clark
- Jeff Alstott
- Jeremy Nixon
- Jingying Yang
- Joshua Kroll
- Lisa Dyer
- Miles Brundage
- Molly Welch
- Paul Christiano
- Peter Eckersley
- Seán Ó hÉigeartaigh
- Shahar Avin
- Shixiong (Austin) Zhang
- Teddy Collins
- Tim Hwang
- William Isaac

---

[82]Our intent with this section of the report is to be transparent with readers about our process and to acknowledge some of the voices and methods missing from that process. We also hope that providing information about our process could be helpful for those considering similar multi-stakeholder research projects.

Given our initial focus on synthesizing and extending existing work, we brought together experts in dimensions of trust that were identified in a pre-workshop white paper, several of which are discussed in this report (such as secure enclaves, third party auditing, and privacy-preserving machine learning). However, a number of voices were missing from that conversation. The workshop could have benefited in particular from greater gender diversity (fewer than one third of participants were women, and none were members of trans or non-binary communities); greater racial diversity (people of color and especially women of color were under-represented, particularly given the number of women of color with relevant expertise on trust in AI development); greater representation of low income communities; greater representation of people with disabilities; and greater geographic, political, philosophical, and religious diversity.

Following the workshop, the corresponding authors led a multi-stakeholder writing, editing, and feedback process. A subset of workshop attendees opted to engage in the writing process for the report. After the first round of writing and editing, we tried to incorporate new authors with complementary expertise to those at the original workshop. Notwithstanding these efforts, not all dimensions of trust in AI development (or even verifiable claims in AI development) were represented in the expertise of the authors. As such, over time and especially in response to external reviewer feedback, we progressively narrowed the scope of the report in order to avoid overreach and "stay in our lane" topic-wise. One such shift was a move from discussing trust in AI development generally to verifiable claims specifically as the focus of the report.

The report was written in a semi-modular fashion. Experts in particular areas drafted subsections on mechanisms or research areas with which they are familiar. These subsections were revised substantially over time in response to author and reviewer feedback, and to shifts in the framing of the report. External reviewers provided feedback on specific portions of the report for clarity and accuracy, although the report as a whole was not formally peer reviewed.

Given the number of authors involved and the wide-ranging nature of the report, it was difficult to ensure that all authors were fully supportive of all content throughout the writing process. Where appropriate, we have included footnotes to clarify process and attribution.

II Key Terms and Concepts

AI: we define artificial intelligence (AI) as any digital system capable of performing tasks commonly thought to require intelligence, with these tasks often being learned from data and/or experience.[83]

AI system: we define an AI system as a *software* process (with the characteristics of AI mentioned above), running on physical *hardware,* under the direction of humans operating in some *institutional* context. This framing of AI systems informs the discussion of mechanisms in the report. The properties of the software, hardware, and institutions at work in a given AI system are all potentially relevant to the verifiability of claims made by an AI developer. Focusing on any of these to the exclusion of others could result in a flawed understanding of the overall system.

AI development: we use the term AI development to refer to the process of researching, designing, testing, deploying, or monitoring AI as defined above.

AI developer: we use the term AI developer to refer to individuals or organizations involved in AI development as defined broadly above, including research scientists, data engineers, and project managers at companies building AI-based products and services as well as those in analogous roles in academia, government, or civil society. Given the major role played by technology companies in contemporary AI development, we pay particular attention to such companies in the report, while recognizing that different contexts will require different approaches to earning trust.

Responsible AI development: we follow Askell et al. in defining responsible AI development as follows [1]:

> "Responsible AI development involves taking steps to ensure that AI systems have an acceptably low risk of harming their users or society and, ideally, to increase their likelihood of being socially beneficial. This involves testing the safety and security of systems during development, evaluating the potential social impact of the systems prior to release, being willing to abandon research projects that fail to meet a high bar of safety, and being willing to delay the release of a system until it has been established that it does not pose a risk to consumers or the public."

Transparency: we define transparency as making information about the characteristics of an AI developer's operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.[84] Transparency can benefit from the open publication of AI systems (including code, data, and models), though privacy, safety, and competitive considerations prevent this from being appropriate in all cases.[85]

---

[83]Some distinctions are made between different phases of AI development in the report, although the authors have also found it helpful to take a broad view of look at AI development: in many cases, the same mechanisms (especially institutional ones) are applicable to multiple phases of development, and *AI development* was found to be the most appropriate catch-all term for such purposes. A recent and representative example of a more granular breakdown, from Toreini et al., distinguishes data-related steps (data collection, data preparation, and feature extraction) from model-related steps (training, testing, and inference) [129]. In practice, these steps are not followed in a linear manner, since (e.g.) testing may inform changes to the training process and data may be improved over time.

[84]See, for example, the Fairness, Accountability, and Transparency in Machine Learning (FATML) workshop and community, which was followed by the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).

[85]The necessity and sufficiency of transparency as an ideal for technical systems has also been critiqued in recent years, such as from Ananny and Crawford [130] and Kroll et. al. [131].

Realizing transparency in AI development requires attention to institutional mechanisms and legal structures, particularly when scrutinizing developers' incentives and scope of activities.

Trust and **trustworthiness**: these concepts have been extensively explored by researchers, though a consensus account across domains remains elusive. Substantial prior and ongoing work focuses on how these concepts manifest in AI development. This includes academic work [129],[86] government-associated efforts such as the European Union's High-Level Expert Group on AI [133], and industry efforts [134] [135]. Our report focuses on a particular subset of what these concepts entail in the practice of AI development, namely the verifiability of claims about AI development, and more specifically, the verifiability of claims about safety, security, privacy, and fairness.

Several common frameworks for thinking about trust suggest a premium on the verifiability of claims even when they do not reference these terms explicitly. For example, Mayer et al.'s widely cited work [136] identifies benevolence, integrity, and ability as three pillars of trustworthiness. In the context of AI, a developer might make claims that suggest their pursuit of a benevolent goal (e.g., by adopting a set of ethical principles), but this needs to be accompanied by the skills and resources (ability) to achieve that goal as well as sufficient transparency and incentives to ensure consistent follow-through (integrity).

Another prominent definition of trust which supports a focus on verifiable claims comes from Gambetta (paraphrased below):[87]

> "When we say we trust someone or that someone is trustworthy, we implicitly mean that we assess that the probability [they] will take actions that are beneficial (or at least not detrimental) is high enough for us to consider engaging in some form of cooperation with [them]."

Here, too, the ability to scrutinize the claims and commitments made by an AI developer can provide calibration regarding the extent to which trust is appropriate in a given context.

The verifiability of claims is also a key theme in the study of trust in the context of international relations and arms control [138]. Ronald Reagan's famous "trust but verify" (a proverb taught to him by advisor on Russian affairs Suzanne Massie [139]) emphasized the value of generating and assessing evidence of compliance with arms control agreements between the United States and Soviet Union. AI, verification, and arms control are discussed further in **Appendix IV**.

This report is not intended to make novel contributions to the theory of trust or trustworthiness, but rather to explore verifiable claims as a building block of trustworthy AI development. When we use terms such as "earn trust" or "calibrate trust" in the report, these are meant to refer to cases in which evidence is provided to substantiate claims about an actor's behavior (including AI development).

---

[86]See [132] for a useful literature review.
[87]This is a modification of Gambetta's definition [137].

III   The Nature and Importance of Verifiable Claims

Verifiable[88] **claims** are statements for which evidence and arguments can be brought to bear on the likelihood of those claims being true. Verifiable claims are sufficiently precise to be falsifiable, and the degree of attainable certainty in such claims will vary across contexts.

AI developers regularly make claims regarding the properties of AI systems they develop as well as their associated societal consequences. Claims related to AI development might include, e.g.:

- We will adhere to the data usage protocols we have specified;

- The cloud services on which our AI systems run are secure;

- We will evaluate risks and benefits of publishing AI systems in partnership with appropriately qualified third parties;

- We will not create or sell AI systems that are intended to cause harm;

- We will assess and report any harmful societal impacts of AI systems that we build; and

- Broadly, we will act in a way that aligns with society's interests.

The verification of claims about AI development is difficult in part due to the inherent complexity and heterogeneity of AI and its supporting infrastructure. The highly dispersed ecosystem means there are many actors and sets of incentives to keep track of and coordinate. Further, the speed of development also means there is less time for claims to be carefully expressed, defended, and evaluated. And, perhaps most critically, claims about AI development are often too vague to be assessed with the limited information publicly made available.

Notwithstanding these challenges, there are at least three distinct reasons why it is highly desirable for claims made about AI development to be verifiable.

First, those potentially affected by AI development–as well as those seeking to represent those parties' interests via government or civil society–deserve to be able to scrutinize the claims made by AI developers in order to reduce risk of harm or foregone benefit.

Second, to the extent that claims become verifiable, various actors such as civil society, policymakers, and users can raise their standards for what constitutes responsible AI development. This, in turn, can improve societal outcomes associated with the field as a whole.

Third, a lack of verifiable claims in AI development could foster or worsen a "race to the bottom" in AI development, whereby developers seek to gain a competitive edge even when this trades off against important societal values such as safety, security, privacy, or fairness [1]. In both commercial (e.g.,

---

[88]While the report does discuss the technical area of formal verification at several points, the sense in which we use "verifiable" is distinct from how the term is used in that context. Unless otherwise specified by the use of the adjective "formal" or other context, this report uses the word verification in a looser sense. Formal verification seeks mathematical proof that a certain technical claim is true with certainty (subject to certain assumptions). In contrast, this report largely focuses on claims that are unlikely to be demonstrated with absolute certainty, but which can be shown likely or unlikely to be true, i.e. trustworthy, or untrustworthy through relevant arguments and evidence.

autonomous vehicles) and non-commercial (e.g., military) contexts, verifiable claims may be needed to foster cooperation rather than race-like behavior.

Without the ability to verify AI-related claims, the decision of how and whether to interact with AI systems must be made without information that could bear on the desirability of having that interaction. Given the large (and growing) stakes of AI development, such an information deficit is ethically untenable. An environment of largely unverifiable claims about AI could encourage extreme reactions to AI in particular situations (i.e., blind trust or blind rejection), resulting in both over-use and under-use of AI. The world instead needs trust in AI development to be well-calibrated, i.e. it should be the case that confidence in certain claims or actors is proportional to the available evidence. The benefits and risks of AI are many, and need to be appraised with context and nuance.

In the past decade, claim-oriented approaches have been developed in order to structure arguments about the safety of engineered systems [140], and we draw inspiration from such approaches in this report. One result of such work is the introduction and standardization of assurance cases in numerous domains. An assurance case is a documented body of evidence that provides a convincing and valid argument regarding a top-level claim (such as the safety of a nuclear power plant), and presents a structured justification in support of that claim to decide the status of it. Assurance cases are often required as part of a regulatory process (e.g., a certificate of safety being granted only when the regulator is satisfied by the argument presented in a safety case).[89]

This work matured into the widely-used Claims, Arguments, and Evidence (CAE) framework.[90] CAE is often the framework of choice in aviation, nuclear, and defense industries worldwide to reason about safety, security, reliability and dependability, and recent work has begun applying CAE to the safety analysis of AI systems.[91]

The CAE framework consists of three key elements. **Claims** are assertions put forward for general acceptance. They're typically statements about a property of the system or some subsystem. Claims asserted as true without justification are assumptions, and claims supporting an argument are subclaims. **Arguments** link evidence to a claim, which can be deterministic, probabilistic, or qualitative.[92] They consist of "statements indicating the general ways of arguing being applied in a particular case and implicitly relied on and whose trustworthiness is well established" [144], together with validation of any scientific laws used. In an engineering context, arguments should be explicit. **Evidence** serves as the basis for justification of a claim. Sources of evidence can include the design, the development process, prior experience, testing, or formal analysis.

For a sufficiently complex AI system or development process, a wide variety of mechanisms will likely

---

[89]Assurance cases are primarily concerned with demonstrating the validity (or otherwise) of the resulting argument and have two main roles: logical reasoning and communication. Cases are usually integrated within a regulatory process that provides for independent challenge and review. There can be a number of stakeholders including public officials, developers, certifiers, regulators. Communication is thus essential to create a shared understanding between the different stakeholders, build confidence and consensus

[90]See, e.g., this discussion of CAE in the nuclear safety context [141] and Uber's use of GSN [142]

[91]See Zhao et al. [143].

[92]Arguments are presented in a form of defeasible reasoning of top-level claims, supported by the available evidence; and driven by practical concerns of achieving the required goal in the best possible way considering the existing uncertainties, point of views, concerns and perspectives of different stakeholders. Such argumentation is expected to be multidisciplinary, and cover a wide range of mechanisms, which we aim to address. To support CAE, a graphical notation can be used to describe the interrelationship of claims, arguments, and evidence. Claim justifications can be constructed using argument blocks–concretion, substitution, decomposition, calculation, and evidence incorporationâĂŤas well as narrative analyses that describe the claims, arguments, and evidence in detail.

need to be brought to bear in order to adequately substantiate a high-level claim such as "this system was developed in accordance with our organization's ethical principles and relevant laws."

IV AI, Verification, and Arms Control

At an international level, arms control is a possible approach to addressing some of the risks of AI development in a military context. Arms control involves similar issues to those discussed earlier (namely, the need for credible commitments and close attention to transparency and incentives) in non-military contexts. In this subsection, we provide an overview of the relationship between verifiable claims and arms control applied to AI.

Arms control is a special case of regulation, in which nation-states cooperate to self-regulate weapons technologies under particularly challenging conditions [145] [146] [147] [148] [149]. Unlike in domestic regulation, there is no external actor to force compliance if states violate the terms of an arms control agreement. Instead, states generally rely on reciprocity to enforce arms control agreements. If states violate an agreement, they can often expect others to follow suit and develop the weapon themselves.

Formal agreements such as treaties act as coordination mechanisms for states to reach agreement, but do not directly enforce compliance. Some treaties include verification regimes to help increase visibility among states as to whether or not others are complying with an agreement, as a means of facilitating trust [150] [151]. But it is on states themselves to take action if others are found violating an agreement, whether through sanctions, reciprocal weapons development, military action, or other tools of statecraft.

Arms control is inherently challenging not only because there is no third-party enforcement mechanism, but because states may be incentivized to violate agreements if they believe that doing so may give them an edge against competitors [152] [153] [154]. This tension is exacerbated if it is challenging to verify other states' behavior. States may assume others are cheating and developing a prohibited technology in secret, incentivizing them to do so as well or risk falling behind a competitor. Arms control agreements can also be more challenging to hold together if a technology is more widely accessible to a larger number of actors and if defection by one actor generates incentives for others to defect. There are many cases in history in which nation-states genuinely desired mutual restraint for certain weapons, such as turn of the century rules regulating submarines, air-delivered weapons, and poison gas, but states were unable to achieve effective cooperation in wartime for a variety of reasons.

Despite these hurdles, there have been successful examples of arms control for a number of weapons, including: chemical [155] [156] and biological [157] weapons; land mines [158] [159]; cluster munitions [160]; blinding lasers [161]; exploding bullets; limits on the proliferation, quantity, and deployment of nuclear weapons [162]; anti-ballistic missile systems [163]; weapons of mass destruction in space; and weapons on the Moon or in Antarctica. There are also examples of mutual restraint with some weapons despite the lack of formal agreements, including neutron bombs, kinetic (debris-causing) anti-satellite weapons, and certain forms of bayonets.

Even these successes highlight the limitations of arms control, however. Some treaties have collapsed over time as more nations gained access to the underlying technology and did not abide by the prohibition. And even the most successful prohibitions, such as those on chemical and biological weapons, have failed to rein in rogue regimes or terrorists. Despite the widespread global condemnation of chemical weapons, Bashar al Assad has used them in Syria to murder civilians, with minimal consequences from the international community.

In general, arms control is more likely to succeed when:[93] (1) there are clear lines between which

---

[93]See, e.g., Crootof [164], Watts [165], and Scharre [166].

weapons are prohibited and which are permitted; (2) the perceived horribleness of a weapon outweighs its military value; (3) states have the ability, either through formal verification regimes or other mechanisms, to ensure that others are complying with the regulation; and (4) fewer states are needed for an agreement to work. Regulation can occur at multiple points of technology development, limiting or prohibiting access to the underlying technology, weapons development, production, and/or use. Note also that while some of the variables above are exogenous from the perspective of the AI community, others are potentially amenable to influence (e.g., research could potentially improve the distinguishability of offensive and defensive uses, or improve the general traceability and interpretability of AI systems).

AI will likely be a transformative technology in warfare [167]. Anticipating this transformation, many in the scientific community have called for restrictions or bans on military AI applications. Because AI is a general-purpose enabling technology with many applications, broad bans on AI overall are unlikely to succeed. However, prohibitions on specific military applications of AI could succeed, provided states could agree to such limits (requiring that the terms be compatible with the incentives of each party) and that appropriate means of verifying compliance are developed and implemented.

AI technology has certain attributes that may make successful restraint challenging, however. These include its widespread availability, dual use or "omni-use" nature [168], the difficulty in drawing clear lines between acceptable and unacceptable AI applications, and the challenges of verifying compliance, which are at least as difficult as those found in non-military contexts and perhaps more challenging given the more adversarial context.

One special case worth highlighting is the development of lethal autonomous weapon systems (LAWS). An international LAWS ban has been the subject of discussion at the UN Convention on Certain Conventional Weapons (CCW) since 2014. There are many arguments made in support of restrictions on LAWS. Three relevant arguments are: (1) their use is immoral because AI systems will not in the foreseeable future understand the moral, psychological, and social context at the time of killing a person (unlike a human, who could decide to not press the trigger) [169] [170]; (2) the state of the technology today would preclude their use under international law in anything but isolated cases, such as undersea where civilians are not present; and (3) they might proliferate easily, enabling misuse [168] [171]. Those skeptical of a ban on lethal autonomous weapon systems often reference mutual distrust as a reason for development: "if we don't develop them, others will, putting us at a strategic disadvantage" is a refrain echoed by several great powers.

Avoiding such an impasse requires grappling with the issue of trust head-on, and closely attending to the complexities of AI development in practice. Similar to how AI ethics principles need to be supplemented with mechanisms that demonstrate the implementation of such principles, trust in military-relevant AI systems must be supported by mechanisms based on a rigorous analysis of the dynamics of military AI development. Lethal autonomous weapons are currently the focus of much related discussion, though the use of AI in cyberwarfare and nuclear command and control have also been highlighted as challenging areas in recent years. Some early work in the direction of coordination on AI among great powers [172] has called attention to the need for early dialogue on AI safety and security. Other work has fleshed out the notion of meaningful human control as a cornerstone of lethal autonomous weapon system governance [173] [161] [174].

The AI community and advocates in other disciplines have played a key role in bringing this issue to the attention of the international community [175] [176] [177] [178] [179] [180] [181]. Similar efforts by expert communities have improved prospects for arms control in prior contexts such as nuclear weapons [182] [183]. There remains more to be done to raise the profile of the issue among policymakers, and

to identify appropriate steps that individuals and organizations in the AI community can take to forestall the development of potentially harmful systems.

AI researchers could contribute technical expertise that helps identify potential governance mechanisms in this context. For example, AI researchers, working with arms control experts in other disciplines, could scrutinize proposals such as defensively-oriented AI weapons systems that could target lethal autonomous weapons (but not humans) and help think through different means of limiting proliferation and ensuring human accountability. The AI community's distributed expertise in the process of AI development and the feasibility of different technical scenarios could thus be brought to bear to limit AI "arms racing" and prevent a race to the bottom with respect to the safety, security, and human-accountability of deployed military AI systems [166]. The feasibility of verifying, interpreting, and testing potential AI systems designed for various purposes, as well as the feasibility of using different points of control for governance of the supply chain (e.g., the computing and non-computing hardware associated with autonomous weapons vs. the underlying software), are all issues to which AI expertise is relevant.

Of the various inputs into AI development (including hardware, software, data, and human effort), it's worth noting that hardware is uniquely governable, at least in principle. Computing chips, no matter how fast, can perform only a finite and known number of operations per second, and each one has to be produced using physical materials that are countable, trackable, and inspectable.[94] Similarly, physical robots rely on supply chains and materials that are in principle trackable. Computing power and hardware platforms for robotics are thus potentially amenable to some governance tools used in other domains that revolve around tracking of physical goods (e.g., export controls and on-site inspections).

While it is unclear what hardware-based verification efforts might look like in the context of AI-related arms control, and how feasible they would be, one might imagine, e.g., a bottleneck in the manufacturing process for lethal autonomous weapons. In contrast to such a bottleneck, AI-related insights, data, code, and models can be reproduced and distributed at negligible marginal cost, making it inherently difficult to control their spread or to use them as a metric for gauging the capabilities of an organization with respect to developing lethal decision-making systems.[95] Given such considerations, it is incumbent upon stakeholders to consider ways in which the distinctive properties of hardware might be leveraged in service of verifying any future arms control agreements.

---

[94]We emphasize that this discussion is exploratory in nature, and that there would be major practical challenges involved in acting on these high-level ideas. Our goal in highlighting the unique affordances of hardware is to foster creative thinking about these issues rather than to suggest that there is a readily available solution to the weaponization of AI.

[95]Another potential approach would be to impose constraints on the physical characteristics of AI-enabled military systems, such as their range, payload, endurance, or other non-AI related physical attributes.

V Cooperation and Antitrust Laws

Collaborations between competing AI labs, even for beneficial purposes such as enabling verifiable claims, can raise antitrust issues. Antitrust law is also known as "competition law" or "anti-monopoly law" outside the US. This section primarily addresses US antitrust law, but given the international nature of AI development and markets, attention to the international legal implications of industry collaborations is warranted.

US antitrust law seeks to prevent "unreasonable" restraints on trade [184]. Unreasonableness, in turn, is tested by economic analysis [185]–specifically, a "consumer welfare" test [186]. Although recent academic [187] and popular [188] proposals challenge the wisdom and usefulness of this test, consumer welfare remains the guiding principle for antitrust courts [186].

Antitrust law generally condemns per se particularly harmful restraints on trade,[96] such as direct restraints on price, output levels, competitive bidding, and market allocation [190]. Other practices are analyzed by the Rule of Reason, "according to which the finder of fact must decide whether the questioned practice imposes an unreasonable restraint on competition, taking into account a variety of factors, including specific information about the relevant business, its condition before and after the restraint was imposed, and the restraint's history, nature, and [net] effect." [191] Importantly, courts in the past have consistently **rejected** safety-based (and other public policy-based) defenses of anticompetitive behavior [192].

In a leading case on point, *National Society Professional Engineers v. United States*, the US Supreme Court reasoned that by adopting antitrust laws, Congress had made a "basic policy" decision to protect competition [193]. The Court therefore concluded that the defendant's argument that price competition between engineers was unsafe "[wa]s nothing less than a frontal assault on the basic policy of the [antitrust laws].[97]" "In sum," the Court held, "the Rule of Reason does not support a defense based on the assumption that competition itself is unreasonable."[98]

None of this implies, however, that collaborations between competitors are always anticompetitive and therefore violative of antitrust laws. American antitrust authorities have acknowledged that collaborations between competitors can have important procompetitive benefits, such as enabling new products to be developed, sharing useful know-how, and capitalizing on economies of scale and scope [194]. These benefits need to be balanced against possible harms from collaboration such as reduced competition on pricing or output levels, reducing the pace of progress, or increasing the uniformity of outputs.[99]

If the right antitrust governance procedures are in place, joint activities between competitive AI labs can both enhance consumer welfare and enhance intra-industry trust. Nevertheless, it is important to not allow the goal of supporting verifiable claims to provide cover for practices that would harm consumer welfare and therefore erode trust between society and AI labs collectively.

---

[96]Practices are condemned per se "[o]nce experience with [that] particular kind of restraint enables the Court to predict with confidence that [antitrust analysis] will condemn it..."[189]

[97]*See id*. at 695.

[98]*Id* at 696.

[99]*See generally id.*

VI   Supplemental Mechanism Analysis

A   Formal Verification

Formal verification techniques for ML-based AI systems are still in their infancy. Challenges include:

- Generating formal claims and corresponding proofs regarding the behavior of ML models, given that their output behavior may not always be clear or expected relative to the inputs (e.g., an ML model will not necessarily display the same behavior in the field that it exhibited under a testing environment). As a consequence, traditional formal properties must be reconceived and redeveloped for ML models;

- The difficulty of correctly modeling certain ML systems as mathematical objects, especially if their building blocks cannot be formalised within mathematical domains utilized by existing verification techniques; and

- The size of real-world ML models, which are usually larger than existing verification techniques can work with.

Some preliminary research [195] has attempted to find ways of specifying types of ML robustness that would be amenable to formal verification: for example, pointwise robustness. Pointwise robustness is a property that states that an ML model is robust against some model of adversarial attacks and perturbations at a given point [58]. However, researchers [140] have observed that the maturity and applicability of both the specification and corresponding techniques fall short of justifying functionality, dependability, and security claims. In general, most system dependability properties have gone unspecified,[100] and these methodologies have not accounted for specifications that are more unique to ML-based systems.

Other efforts [69] [196] [197] aim to verify more traditional specifications regarding ML algorithms. Some of these techniques require functional specifications, written as constraints, to be fed into specialized solvers which then attempt to verify that they hold on a constraint model of the ML system. The generalization of these techniques to deep learning is challenging because they require well-defined, mathematically specifiable properties as input which are not unique to ML algorithms (given that such properties do not easily lend themselves to such specifications). These techniques are only applicable to well-specified deterministic or tractable systems that can be implemented using traditional methods (e.g., the C programming language) or via ML models. As a consequence, these techniques cannot be straightforwardly applied to arbitrary contexts, and domain-specific effort is currently required even to specify properties of interest, let alone verify them.

Indeed, there is much progress to be made with regard to the verification of deep neural networks, but formal verification can still be effectively utilised to reinforce non-ML software employed to construct the ML model itself. For example, researchers have demonstrated a methodology in which developers can use an interactive proof assistant to implement their ML system and prove formal theorems that their implementation is free of errors [198]. Others have shown that overflow and underflow errors within supporting software can propagate and affect the functionality of an ML model [199]. Additionally, researchers have identified a number of different run-time errors using a traditional formal methods-based static-analysis tool to analyze YOLO, a commonly used open source ML vision software [140]. Issues identified include:

---

[100]For example: functionality, performance, reliability, availability, security, etc.

- A number of memory leaks, such as files opened and not closed, and temporarily allocated data not freed, leading to unpredictable behavior, crashes, and corrupted data;

- A large number of calls to free where the validity of the returned data is not checked. This could lead to incorrect (but potentially plausible) weights being loaded to the network;

- Potential "divide by zeros" in the training code. This could lead to crashes during online training, if the system were to be used in such a way; and

- Potential floating-point "divide by zeros," some of which were located in the network cost calculation function. As noted above, this could be an issue during online training.

We note that many of the above errors are only applicable to languages such as C and C++ (i.e., statically typed languages), and not Python, a language widely used in the implementation of numerous ML libraries and frameworks. As a dynamically typed language, Python brings about a different set of program errors not typically exhibited by statically typed languages (e.g., type errors). Unfortunately, formal verification techniques for the analysis of Python code are inherently limited, with linters and type checkers being the main available source of static analysis tools.

Though the Python situation differs from that encountered with C and C++, there are many ways that potential faults arising from Python could affect the functionality of an ML model. This is a large gap within the formal verification field that needs to be addressed immediately, given the deployment of safety-critical AI systems, such as autonomous vehicles, utilizing Python. Previous research efforts[101] [200] have attempted to formalise a subset of Python that would be amenable to verification; however, it has been notoriously difficult to formalise and verify [201] dynamically typed languages. Although optional static type hinting is now available for Python,[102] "the Python runtime does not enforce function and variable type annotations. [Hints] can be used by third party tools such as type checkers, IDEs, linters, etc." Furthermore, it is unlikely that the ML community will constrain themselves to subsets of Python which are statically-typed.[103]

Formal verification techniques have been widely deployed for traditional safety-critical systems (as required by IEC 61508) for several decades, and have more recently been adopted by some tech companies for specific applications.[104] However, the rapid introduction of machine learning in these environments has posed a great challenge from both a regulatory and system assurance point of view. The lack of applicable formal verification techniques for AI systems stifles the assurance avenues required to build trust (i.e., regulations, verification, and validation), curbing the potential innovation and benefits to be gained from their deployment. The following open research problems must thus be addressed to allow formal verification to contribute to trust in AI development:

- Creation of specifications unique to AI, with corresponding mathematical frameworks, to contribute to assurance of AI systems;

- Creation of novel formal verification techniques which can address the newly defined specifications mentioned above; and

---

[101]See Python semantics: (https://github.com/kframework/python-semantics).

[102]See official Python documentation (https://docs.python.org/3/library/typing.html) and MyPy (https://github.com/python/mypy).

[103]See the discussion following the feature request in the TensorFlow codebase (https://github.com/tensorflow/tensorflow/issues/12345).

[104]See Infer, an open source static analyzer (https://fbinfer.com/).

- Collaboration between ML and verification researchers resulting in deep learning systems that are more amenable to verification [202].

B Verifiable Data Policies in Distributed Computing Systems

Current IT systems do not provide a mechanism to enforce a data policy (e.g., sharing restrictions, anonymity restrictions) on data that is shared with another party - individuals and organizations are required to trust that the data will be used according to their preferences. Google's Project Oak[105] aims to address this gap, by providing a reference implementation of open source infrastructure for the verifiably secure storage, processing, and exchange of any type of data.

With Oak, data is collected and used as it is today, but it is also accompanied by enforceable policies that define appropriate access, collection, and use. Data is stored in encrypted enclaves and remote attestation between enclaves ensures that only appropriate code ever gets direct access to the secured data (i.e. within the limits of what can be verified and as defined by a configurable policy), and processing of the data creates a verifiable record. In the long term, Google's objective is to provide formal proofs such that core properties of the system can be verified down to the hardware. Platforms that implement this infrastructure could then form the bedrock of all sorts of other services from messaging, machine learning, and identity management to operating system hosting, making meaningful control technically feasible in a way that it is not today.

Oak uses enclaves and formal verification. Taken together, it is possible to verify that data is only processed in a way that complies with a configurable policy that goes with it. In short, data lives in enclaves and moves from one enclave to another only when the sending enclave is able to convince itself that the receiving enclave will obey the policy that goes with the data and will itself perform the same verification step before sending the data (or data derived from it) on to other enclaves. Movement outside enclaves is only permitted when encrypted with keys available only to the enclave, or as allowed by policy (for example, to show the data to specific people or when adequately anonymized, again specified by policy).

Oak combines formal verification and remote attestation with binary transparency. Oak is being developed entirely as an open source project - this is deliberate and necessary. Because Oak is open source, even in the absence of formal proofs, any independent third party (whether an individual researcher, regulatory body, or consumer advocacy group) can examine Oak's source code and confirm that the implementation matches the expected behavior. With the correctness of Oak confirmed insofar as possible, a given Oak virtual machine needs to be able to attest that it is running the "correct" binary of Oak. This attestation makes it possible for the client (or sending enclave) to assure itself that the requesting enclave will follow any policies on the data, because it knows that the Oak policy enforcement system is running on that enclave and is "truly" Oak - that is: matches the binary of Oak known from the open source repository.

Usability and independent auditability are crucial to an Oak system's utility. Four types of actors are expected to interact with Oak:

- End users: people who will use Oak apps;

- Application developers: people who will build Oak apps;

- Policy authors: people who will define and manage the policies that accompany the data in an Oak app;

---

[105]More detail on Oak's technical aspects, including instructions for how to write programs targeting the current iteration of Oak, can be found on the Oak GitHub repo: `https://github.com/project-oak`

- Verifiers: people who will add credibility to an Oak app by verifying that the policy is upheld.

A user-centric design perspective points to many questions such as: What do people need to know about an Oak app when making a decision about whether to use it? How will people understand the effective policy an Oak app is bound by? How will people's preferences be captured? How will we help people find verifiers to delegate trust decisions to? If Oak apps and data policies change, or a verifier's assessment of a policy changes (which we expect can and will happen), how is this change communicated to the people who need to know?

As important as it is to ensure end-users avoid critical mistakes and can understand the impact of Oak, it is even more important to ensure developers are able to avoid critical mistakes when using Oak. This requires deliberate design of how app makers will build, deploy, and debug.

An Oak node without a correct and useful policy is useless. Oak does not provide privacy by default, it does so only if the policies specified are privacy-preserving. Thus, the user experience of specifying and maintaining the policy that accompanies data is crucial to the successful use of Oak. Policy authors will begin with a set of policy goals that will be refined into a natural language representation of a set of rules and will likely be translated into a set of rules that can be enforced by an Oak node. The policy language for those rules will need to be determined based on the types of protection that is relevant to use cases and the rules that can be verifiably enforced. In some cases, there will be desirable attestations that cannot be formally verified (e.g., non-commercial use). Depending on the context, policies may be generated and managed by a group of people, sometimes the developer and sometimes a cross-functional team from within a company.

C Interpretability

What has interpretability research focused on?

Interpretability research includes work in areas such as explaining a specific prediction [203] [204] [205] [206], explaining global model behavior [207] [195] [208] [209] [210], building more interpretable models [211] [212] [208] [213] [214] [215] [216], interactive visualization tools for human exploration [217] [218] [122] [219] [220], and analyzing functional sub-components of neural networks to understand what they are doing [221]. These areas of work are characterized separately below, although several overlap and interact with one another.

Explaining a specific prediction. Many techniques seek to explain a model's prediction on some given input. For example, one might ask which part of the input–for image models, this might take the form of a heatmap over input pixels [222] [223] [224] [225] [226] [227] [228]–or which training examples [203] [204] [205] [206] were responsible for the model's prediction. By examining the model's reasoning on different input instances through these attribution methods, we can better decide whether or not to trust the model: if a model, say, predicts that an image contains a wolf because of the snowy image background and not because of any actual wolf features, then we can extrapolate that it is likely to misclassify future images with snowy backgrounds [92].

Explaining global model behavior. Instead of explaining individual predictions, other techniques aim to construct human-understandable representations of a model's global behavior, so that users can more directly interrogate what a model might do on different inputs rather than having to extrapolate from explanations of previous predictions. Examples include approximating a complex model with a simpler, more interpretable model (like a shallow decision tree) [207] [195] [229] [209] [210]; or characterizing the role and contribution of internal components of a model (e.g., feature visualization or development of geometric invariance) [223] [228] [86] [230] [231] [232] [233] [234] [235] [236].

Current directions in interpretability research

Building more interpretable models. A separate but complementary line of work seeks to build models that are constrained to be interpretable by design (as opposed to training a complex, hard-to-interpret model and then attempting to analyze it post-hoc with one of the above techniques) [211] [212] [208] [213] [214] [215] [216].

Interactive visualization tools for human exploration. A related research theme is the development of tools that allow humans to interact with, modify, and explore an ML system (e.g., dashboards to visualize model predictions and errors [217] [218]; explanations of model representations [219] [220] [237] [238]; or directly interacting with an AI agent [122].[106]).

Software and tools for practitioners. Most interpretability tools that have been developed are best used by ML researchers for interpretability research. A few software packages that are less research-oriented allow novice users to better understand a dataset's distribution and inspect a model interactively;[107] these packages primarily fall under the category of "interactive visualization tools." Moreover,

---

[106]Also see Google Quickdraw (`https://quickdraw.withgoogle.com/`) and Bach doodle (`https://www.google.com/doodles/celebrating-johann-sebastian-bach`).

[107]See What-if tool for ML model exploration (`https://pair-code.github.io/what-if-tool/`) and Facets

most open-sourced code from interpretability research primarily focuses on the method being introduced and rarely include standardized benchmarks and comparisons with related work, with some exceptions.[108] We hope to see more software packages that empower novice users to use interpretability techniques effectively as well as aid researchers by providing standardized benchmarks for comparing methods. Lastly, much work is focused on interpretability at a particular scale (i.e., individual examples vs. dataset distribution); we desire more work at connecting interpretability work along different axes and scales [234].

---

for data exploration (`https://pair-code.github.io/facets/`).

[108]See saliency repository (/urlhttps://github.com/PAIR-code/saliency), InterpretML (`https://github.com/interpretml/interpret`), and TreeInterpreter (`https://github.com/andosa/treeinterpreter`).

List of Recommendations for Reference

Institutional Mechanisms and Recommendations

1. A coalition of stakeholders should create a task force to research options for conducting and funding **third party auditing** of AI systems.

2. 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.

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

4. AI developers should share more information about **AI incidents**, including through collaborative channels.

Software Mechanisms and Recommendations

5. Standards setting bodies should work with academia and industry to develop **audit trail** requirements for safety-critical applications of AI systems.

6. Organizations developing AI and funding bodies should support research into the **interpretability** of AI systems, with a focus on supporting risk assessment and auditing.

7. AI developers should develop, share, and use suites of tools for **privacy-preserving machine learning** that include measures of performance against common standards.

Hardware Mechanisms and Recommendations

8. 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.

9. 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.

10. 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.