Open Problems in Technical AI Governance

Anonymous authors
Paper under double-blind review

Abstract

AI progress is creating a growing range of risks and opportunities, but it is often unclear how they should be navigated. In many cases, the barriers and uncertainties faced are at least partly technical. Technical AI governance, referring to technical analysis and tools for supporting the effective governance of AI, seeks to address such challenges. It can help to (a) identify areas where intervention is needed, (b) assess the efficacy of potential governance actions, and (c) enhance governance options by designing mechanisms for enforcement, incentivization, or compliance. In this paper, we explain what technical AI governance is, outline why it is important, and present a taxonomy and incomplete catalog of its open problems. This paper is intended as a resource for technical researchers or research funders looking to contribute to AI governance.

1 Introduction

The rapid development and adoption of artificial intelligence (AI) systems[1] has prompted a great deal of governance action from the public sector,[2] academia and civil society (Anderljung et al., 2023a; Moës & Ryan, 2023; Barrett et al., 2023), and industry (Anthropic, 2023a; Microsoft, 2023; Dragan et al., 2024; OpenAI, 2024a), with the aim of addressing potential risks while capitalizing on benefits.

However, key decision-makers seeking to govern AI often have insufficient information for identifying the need for intervention and assessing the efficacy of different governance options. Furthermore, the technical tools necessary for successfully implementing governance proposals are often lacking (Reuel et al., 2024a), leaving uncertainty regarding how policies are to be implemented. For example, while the concept of watermarking[3] AI-generated content has gained traction among policymakers (see for example Council of the European Union, 2024; The White House, 2023b; G7 leaders, 2023; Department for Science, Innovation & Technology, 2023), it is unclear whether current methods are sufficient for achieving policymakers' desired outcomes, nor how future-proof such methods will be to improvements in AI capabilities (Zhang et al., 2023; Ghosal et al., 2023). Addressing these and similar issues will require further targeted technical advances.

In this paper, we aim to catalyse work aimed at addressing such challenges and uncertainties through motivating and introducing the field of **technical AI governance (TAIG)**, providing a taxonomy of subareas, and outlining open problems and research questions. Throughout the paper we use the term technical AI governance to refer to *technical analysis and tools for supporting the effective governance of AI*, where 'technical' pertains to the physical sciences, mathematics, engineering, and related fields; and 'AI governance' refers to the processes and structures through which decisions related to AI are made, implemented and

[1] Our understanding of AI systems follows that of (Basdevant et al., 2024), encompassing infrastructure such as compilers, model components such as datasets, code, and weights, as well as UX considerations.

[2] See, for example, (The White House, 2023a; The White House Office of Science and Technology Policy, 2023; Presidency of the Council of the European Union, 2024; Department for Science, Innovation and Technology et al., 2023a; Department for Science, Innovation and Technology & Office for Artificial Intelligence, 2023; Advisory Body on Artificial Intelligence, 2023; European Commission, 2023; 国家互联网信息办公室国家发展和改革委员会教育部等, 2023)

[3] Watermarks are signals placed in output content that are imperceptible to humans, but easily detectable through application of a specific *detector* algorithm.

enforced.[4] AI governance includes research that is both non-technical (e.g. in politics, economics, law, philosophy, etc.) and technical in nature (e.g. in computer science, engineering, and mathematics). Through our use of this definition we simply refer to this latter category of work within the umbrella category of 'AI governance.'

By the above definition, TAIG can contribute to AI governance in a number of ways, such as by identifying opportunities for governance intervention, informing key decisions, and enhancing options for implementation. For example, deployment evaluations that assess the downstream impacts of a system (see Section 3.4) could help identify a need for policy interventions to address these impacts. Alternatively, being able to design models that are robust to malicious modifications (see Section 6.4) could add to the menu of governance options available to prevent downstream misuse.

In particular, we make the following contributions:

- We introduce the field of technical AI governance and motivate the need for work in this area.
- We present a taxonomy of TAIG arranged along two dimensions: *capacities*, which refer to actions such as access and verification that are useful for governance, and *targets*, which refer to key elements in the AI value chain, such as data and models, to which capacities can be applied.
- Finally, we outline open problems within each category of our taxonomy, along with concrete example questions for future research.

Figure 1 provides an overview of the open problem areas, organized according to the taxonomy. We hope that this paper serves as a resource and inspiration for technical researchers aiming to direct their expertise towards policy-relevant topics.

[4]To the extent that this definition of TAIG includes measures for directly increasing the performance, safety, or robustness of AI systems, we only consider such measures for cases in which they support the governance of AI.

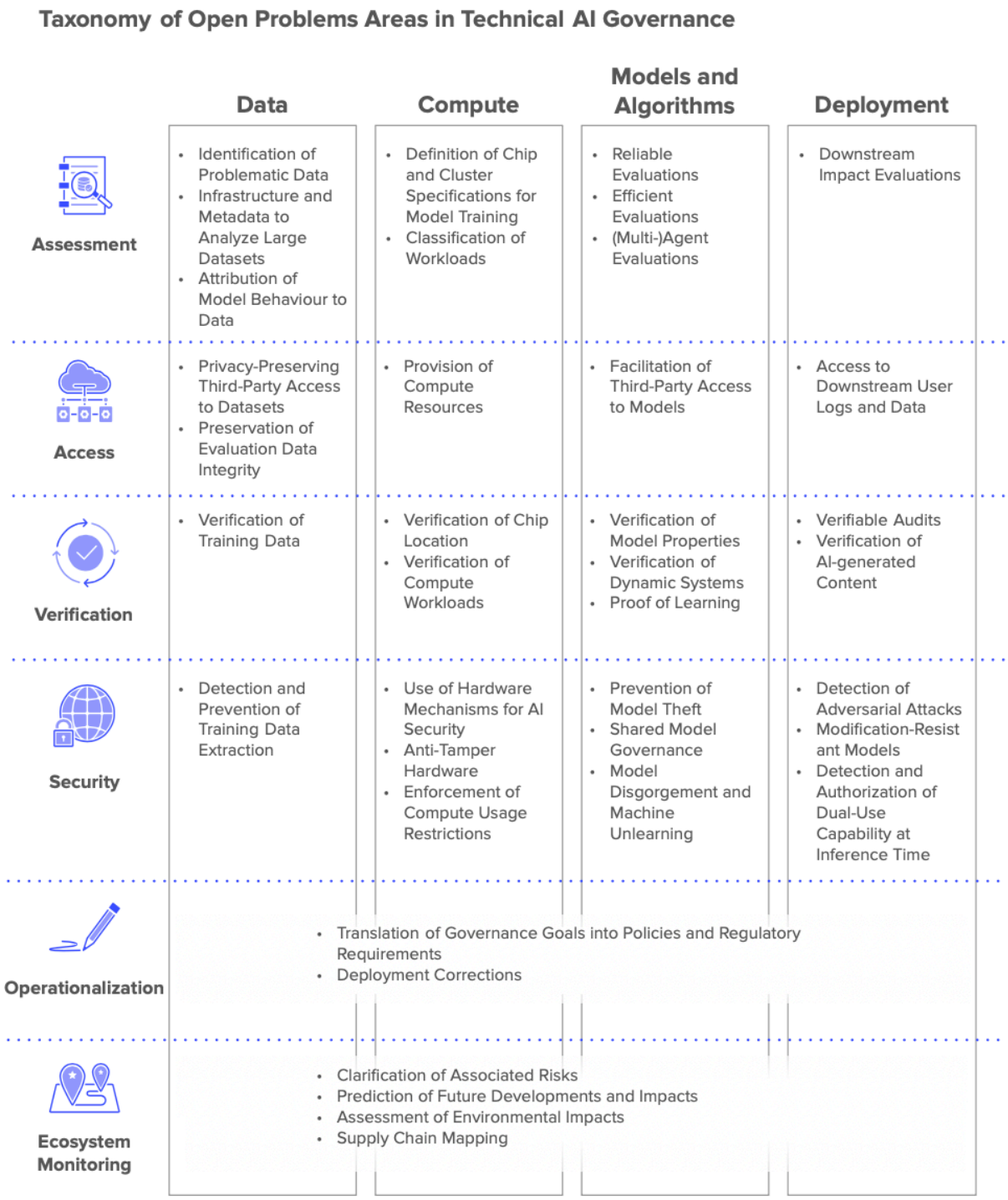


Figure 1: An overview of the open problem areas covered in this report, organized according to our taxonomy.

1.1 Relation to AI Governance

As noted above, we define AI governance as the processes and structures through which decisions related to AI are made, implemented, and enforced. It encompasses the rules, norms, and institutions that shape the behavior of actors in the AI ecosystem, as well as the means by which they are held accountable for their actions.[5] As per our definition above, TAIG consists of *technical analysis and tools for supporting the effective governance of AI*. Here we outline three ways in which TAIG can contribute to AI governance, which we refer to as *identifying*, *informing*, and *enhancing*.[6]

Firstly, **TAIG can identify areas where governance intervention is needed**, through mapping technical aspects of AI systems to social and political concepts, typically conceived of as being addressed through governance. For example, tracking and considering technical advances in AI video generation could allow for more accurate predictions of the risk of video deepfakes, and thus motivate the need for a governance response.

Secondly, **TAIG can inform governance decisions** by providing decision-makers with more accurate information, allowing them to better compare the effectiveness of different governance options. For example, policymakers can choose between different regulatory instruments (for example, registration or disclosure), as well as how they enforce compliance (for example, *ex ante* rules or *post hoc* adjudication), with the efficacy of these options potentially depending on technical details. Information could stem from implemented TAIG methods, such as the outcome of assessments (see Section 3), or TAIG research that maps or monitors the AI ecosystem (see Section 8). For example, more developed risk models for assessing potential harms of AI could inform targeted policies for their mitigation.

Finally, **TAIG can enhance governance options** by providing or enabling mechanisms for enforcing, incentivizing, or complying with mandated requirements. For instance, developing methods for the robust evaluation of models with black-box access could facilitate more comprehensive third-party auditing, thereby enhancing enforcement of safety requirements.

Previous overviews of AI governance, such as Dafoe (2018)'s AI governance research agenda, focus primarily on socio-political governance challenges, including labor displacement, inequality, shifts in national power, and AI race dynamics, or taxonomizes themes, gaps, and issues in AI governance more broadly (Birkstedt et al., 2023). This literature sets the stage for why technical solutions are needed (e.g., an "AI race that sacrifices safety" (Dafoe, 2018) could imply a need for governance-oriented technical security mechanisms). More recent policy documents (see e.g., OECD, 2019; G7, 2023) enumerate high-level issues like privacy concerns, bias, a lack of accountability, and security in AI, while including technical, socio-political, and institutional considerations. Our work contributes by focusing more narrowly on the technical dimension of those issues. For instance, where such reports highlight AI accountability as a governance challenge, our taxonomy offers concrete technical problem areas (like verifiable auditing and data provenance tracking) that would help achieve accountability in practice.

The AI risk literature categorizes both realized harms and anticipated risks of AI (see e.g., Lobel, 2024; Shelby et al., 2023; Raji et al., 2022a; Weidinger et al., 2022; Critch & Russell, 2023). For example, Shelby et al. (2023) presented a human-centered taxonomy of sociotechnical harms from algorithmic systems, identifying five harm categories (representational, allocative, quality-of-service, interpersonal, and societal harms) based on a review of 172 research papers. Similarly, Uuk et al. (2024) conducted a systematic literature review to derive 13 categories of systemic AI risks, including environmental harms, structural discrimination, and loss of control. Similar to the literature on challenges in AI governance, this prior work is relevant to our efforts in that it motivates work on AI governance – indeed, many of the open problems we list (such as developing robust evaluation methods) are motivated by those very risks. By situating our work among these prior catalogs, we underscore that our contribution is a focused extension: we zoom in on the technical layer of AI governance to provide a structured overview of open technical problems that underlie or intersect with identified AI governance challenges.

[5] For other proposed definitions of AI governance, see (Bullock et al., 2022; Daly et al., 2021; Dafoe, 2018).

[6] A useful parallel to TAIG may be the concepts of regulatory technology (RegTech) and supervisory technology (SupTech) in the financial sector (Bank for International Settlements, 2021), which aim to support financial regulation and oversight.

1.2 Scope and Limitations

This paper aims to give a broad overview of open technical problems for AI governance, identifying gaps in existing or suggested governance proposals, while avoiding taking a normative position on their desirability or efficacy. Indeed, the governance aims motivating some of the open questions outlined below may be in tension with each other, and we do not expect their solutions all to be used within the same governance framework. For example, broad access to some AI systems may be in conflict with ensuring their security.

At the same time, we are conscious of the potential pitfalls of techno-solutionism – that is, relying solely on proposed technical fixes to complex and often normative social problems – including a lack of democratic oversight and introducing further problems to be fixed (Michael et al., 2020; Lindgren & Dignum, 2023; Angel & Boyd, 2024; Allen, 2024). Many of the TAIG tools presented below are hypothetical and speculative, and we make no claims about the feasibility of developing solutions. Furthermore, some of the TAIG measures highlighted are dual-use. For example, while hardware-enabled mechanisms for monitoring advanced compute hardware could provide increased visibility into the private development of the largest models, they could also potentially be applied to unreasonably surveil individuals using such hardware for legitimate purposes.

Thus, having solutions to all open problems outlined in this paper will not have *solved* AI governance. On the contrary, careful management will be necessary to determine a balance between capacities that are in tension with each other, and to ensure that dual-use capacities are not misused. Furthermore, many AI governance problems may rely predominantly on non-technical solutions, such as ensuring the appropriate inclusion of countries impacted by AI in international AI governance decision-making (Trager et al., 2023). However, we argue that making progress on the technical problems outlined below can help to ensure more robust AI governance on net.

We view TAIG as related to (and in some places overlapping with) existing areas of research. Namely, TAIG differs from topics in AI safety and alignment in that it is not aimed at directly improving the safety of AI systems. Furthermore, while sociotechnical approaches to AI safety and governance (see e.g. Lazar & Nelson, 2023; Bogen & Winecoff, 2024; Oduro & Kneese, 2024) view "society and technology together as one coherent system" (Chen & Metcalf, 2024), TAIG considers the instrumental value of technical work for enacting governance. Taken together with TAIG, these diverse approaches can serve as complementary methods for mitigating risks and promoting beneficial outcomes of AI (Narayanan et al., 2023). Finally, and as noted above, we view TAIG as a subset of AI governance more broadly, which also includes many topics in diverse disciplines including political science, economics, and philosophy.

We consider some notable fields, topics, and problems to be out of scope for this paper. In particular, technical work that directly improves the performance, safety, or robustness of AI systems, or addresses related ethical concerns – while highly relevant to AI governance – is considered out of scope. Topics regarding government or public-sector use of AI (Margetts & Dorobantu, 2019; Aitken et al., 2022; Margetts, 2022; Straub et al., 2023) or ways in which AI could itself be used to defend against or ameliorate downstream harms of AI (Bernardi et al., 2024) are also out of scope.

1.3 Reader Guide

Table 1: Relevant problem areas organized by reader background

ML Theory	Assessment	3.1.2; 3.1.3; 3.2.1; 3.2.2; 3.3.1; 3.3.2; 3.4.1
	Access	4.1.1; 4.2.1; 4.3.1
	Verification	5.1.1; 5.2.2; 5.3.1; 5.3.3; 5.4.1; 5.4.2
	Security	6.1.1; 6.3.1; 6.3.2; 6.3.3; 6.4.1; 6.4.2; 6.4.3
	Operationalization	7.2
Applied ML	Assessment	3.1.2; 3.3.1; 3.3.2; 3.4.1
	Access	4.3.1; 4.4.1
	Security	6.4.3
	Operationalization	7.1; 7.2
	Ecosystem Monitoring	8.1; 8.2; 8.3
Cybersecurity	Verification	5.2.2
	Security	6.2.1; 6.2.3; 6.3.1; 6.4.3
	Operationalization	7.2
Cryptography	Assessment	3.1.1; 3.2.2
	Access	4.1.1; 4.1.2; 4.2.1; 4.3.1; 4.4.1
	Verification	5.1.1; 5.2.1; 5.2.2; 5.3.3; 5.4.1; 5.4.2
	Security	6.2.1; 6.2.3; 6.3.2; 6.4.3
Hardware Engineering	Assessment	3.1.2; 3.2.1; 3.2.2
	Access	4.2.1
	Verification	5.2.1; 5.2.2
	Security	6.2.1; 6.2.2; 6.2.3; 6.3.1; 6.3.2
Software Engineering	Assessment	3.1.1; 3.1.2; 3.3.2; 3.4.1
	Access	4.2.1
	Verification	5.2.2
	Security	6.2.1; 6.3.1
Mathematics and Statistics	Assessment	3.1.2; 3.4.1
	Ecosystem Monitoring	8.2; 8.3

This paper provides a broad overview of open problems across the taxonomy defined in Section 2. Given the extensive nature of the main content (Sections 3-8), we have structured it for selective reading:

- **Each section is self-contained**, allowing readers to focus on their area(s) of interest.
- Each section begins with a **summary table of problem areas**.
- **Specific open problems** within each area are highlighted in **bold**.
- **Example research questions** are provided in boxes at the start of each subsection.
- Table 1 offers suggested relevant problem areas based on **reader expertise**.
- We attach a two-page **policy brief** in appendix A.

This structure aims to facilitate quick identification of key issues and relevant problems for readers across various backgrounds and interests.

2 Taxonomy

The paper is organized according to a two-dimensional taxonomy, based around capacities and targets. *Capacities* encompass a comprehensive suite of abilities and mechanisms that enable stakeholders to understand and shape the development, deployment, and use of AI, such as by assessing or verifying system properties. These capacities are neither mutually exclusive nor collectively exhaustive, but they do capture what we believe are the most important clusters of technical AI governance. We list all considered capacities, along with descriptions, in Table 2. An overview of our methodology, and in particular how we derived this taxonomy, can be found in Appendix B.

Table 2: Overview of capacities and their importance for AI governance

Capacity	Description	Why it matters for governance
Assessment	The ability to evaluate AI systems, involving both technical analyses and consideration of broader societal impacts.	Enables the identification and understanding of system capabilities and risks, allowing for more targeted governance intervention.
Access	The ability to interact with AI systems, including model internals, as well as obtain relevant data and information while avoiding unacceptable privacy costs.	Enables external research and assessment of AI systems, and aids in fairly distributing benefits of AI across society.
Verification	The ability of developers or third parties to verify claims made about AI systems' development, behaviors, capabilities, and safety.	Establishes trust in AI systems and confirms compliance with regulatory requirements.
Security	The development and implementation of measures to protect AI system components from unauthorized access, use, or tampering.	Ensures the integrity, confidentiality, and availability of AI systems and guards against misuse.
Operationalization	The translation of ethical principles, legal requirements, and governance objectives into concrete technical strategies, procedures, or standards.	Bridges the gap between abstract principles and practical implementation of regulatory requirements.
Ecosystem Monitoring	Understanding and studying the evolving landscape of AI development and application, and associated impacts.	Enables informed decision-making, anticipation of future challenges, and identification of key leverage points for effective governance interventions.

The second axis of our taxonomy pertains to the *targets* that encapsulate the essential building blocks and operational elements of AI systems[7] that governance efforts may aim to influence or manage. They are adapted from categories introduced in (Bommasani et al., 2023b). Each capacity given above can be applied to each target. We structure our paper around the resulting pairs of capacities and targets, with the exception of *operationalization* and *ecosystem monitoring* which cut across all targets. The targets considered in this report are summarized in Table 3.

We recognize that organizational processes undertaken during the development and deployment of AI systems intersect with and shape these targets, and could be considered as regulatory targets in their own right.

[7] (Repeat of footnote 1) Our understanding of AI systems follows that of (Basdevant et al., 2024), encompassing infrastructure such as compilers, model components such as datasets, code, and weights, as well as UX considerations.

Table 3: Overview of targets

Target	Description
Data	The pretraining, fine-tuning, retrieval, and evaluation datasets on which AI systems are trained and benchmarked.
Compute	Computational and hardware resources required to develop and deploy AI systems.
Models and Algorithms	Core components of AI systems, consisting of software for training and inference, their theoretical underpinnings, model architectures, and learned parameters.
Deployment	The use of AI systems in real-world settings, including user interactions, and the resulting outputs, actions, and impacts.

However, we have chosen not to include them as explicit targets in our taxonomy as processes mostly involve non-technical challenges that fall outside the scope of our paper. In cases where processes do face challenges, we address such issues within the context of the most relevant target. For example, compliance with content-creators' right to opt-out is dependent on identifying copyrighted samples in datasets (Section 3.1).

3 Assessment

Evaluations and assessments of the capabilities and risks of AI systems have been proposed as a key component in AI governance regimes. For example, *model evaluations and red-teaming*[8] comprised a key part of the voluntary commitments agreed between labs and the UK government at the Bletchley Summit (Department for Science, Innovation & Technology, 2023). Furthermore, the White House Executive Order on Artificial Intelligence requires developers of the most compute-intensive models to share the results of all red-team tests of their model with the federal government (The White House, 2023a).

The purpose of assessment is to detect problematic behavior or impacts of AI systems before resulting harms can materialize, as well as to ensure systems are safe, robust, and non-discriminatory. However, the assessment of some targets, especially in the context of foundation models, is currently more an art than a science, with a significant number of open challenges (Chang et al., 2024; Weidinger et al., 2023). These issues are exacerbated by the fact that evaluations are expensive to conduct at scale. While assessment and evaluation standards are emerging (National Institute of Standards and Technology (NIST), 2023; UK AI Safety Institute, 2024), there are still fundamental open technical problems that need to be addressed to ensure robust and informative assessments.

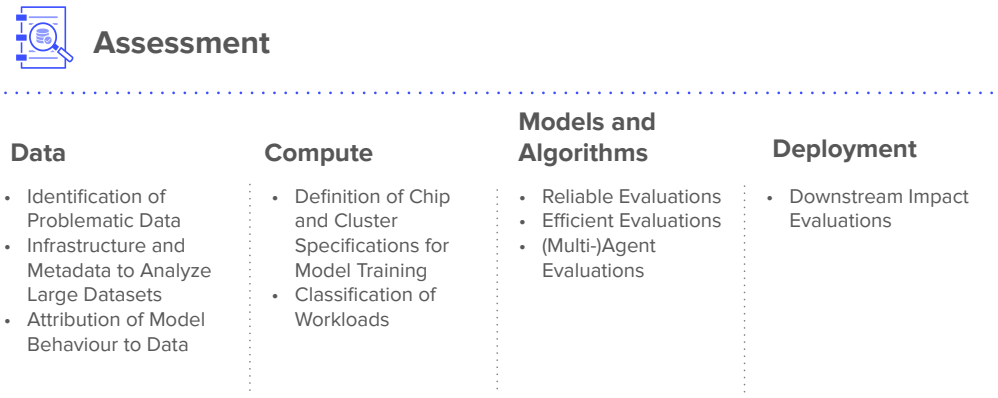


Figure 2: Open problem areas in the *Assessment* capacity, organized by target

3.1 Data

Example Research Questions
1. How can methods for identifying problematic data be scaled to large (on the magnitude of trillions of tokens/samples) datasets? (3.1.1)
2. How can license collection be automated to prevent training on unlicensed data? (3.1.1)
3. How can the accuracy of licenses be ensured when aggregating datasets from multiple sources? (3.1.1)
4. How can problematic data be identified without full/direct access to the dataset? (3.1.1)
5. How can contamination of training data with problematic samples be reliably detected? (3.1.1)
6. How can harmful data be removed from a dataset without facilitating its easy identification by malicious actors? (3.1.1)

[8] *Red-teaming* refers to deliberately trying to find ways to make a system behave poorly, produce harmful outputs, or be misused, in order to identify potential risks and vulnerabilities to be addressed.

7. What license and meta-data reporting requirements could assist in responsible data practices? (3.1.2)
8. What infrastructure is needed to enable researchers to audit large datasets? (3.1.2)
9. How can macro-scale dataset properties, such as persistent bias, be identified and measured? (3.1.2)
10. What information about datasets is necessary for determining their suitability for training? (3.1.2)
11. What is the effect of problematic data on downstream system performance? (3.1.3)
12. Can system behaviors and/or properties be accurately attributed to pretraining and/or fine-tuning data samples? (3.1.3)

3.1.1 Identification of Problematic Data

Motivation: Data plays a central role in the development and resulting capabilities of AI systems. Therefore, issues with data can propagate downstream, resulting in undesirable properties of models. We identify two ways in which data can be problematic.

The first is that data samples may violate some legal or ethical principle, simply by virtue of being included in a dataset. For instance, the presence of a sample could constitute a copyright or privacy violation (Brown et al., 2022; Rahman & Santacana, 2023; Subramani et al., 2023; Marcus & Southen, 2024), data poisoning (Biggio et al., 2012; Steinhardt et al., 2017; Wallace et al., 2020; Carlini & Terzis, 2021; Carlini, 2021; Schuster et al., 2021; Carlini et al., 2023a), or be inherently harmful (Thiel, 2023; Birhane et al., 2021; 2023; Luccioni & Viviano, 2021).

The second way that data could be problematic is if its use in training causes undesirable downstream effects. For instance, models trained on factually incorrect content in training data, such as vaccine disinformation, might replicate those factual inaccuracies. Indeed, (Lin et al., 2022b) demonstrate how models can "[generate] many false answers that mimic popular misconceptions." Alternatively, low-quality data, such as inaccuracies in low-resource languages, can compromise performance of models in those languages (Kreutzer et al., 2022).

Being able to identify problematic data used in the development of AI systems could facilitate various governance measures. For example, developers found to be using inherently harmful data could face penalties for improper use. Alternatively, developers found to be training on data that induces harmful model behaviors could face a higher burden of proof when demonstrating that their models are sufficiently safe for public deployment.

Open problems:

Identifying problematic data given access to the training data. For model developers with full access to the dataset, the major challenge is defining concrete, operationalizable criteria for detecting and removing problematic samples before training. Some problematic samples may be easier to identify than others; for instance, social security numbers can be identified with regexes or direct copies can be easily identified through pattern matching. However, the identification of other forms of problematic data samples poses more of a challenge. For example, understanding whether a data sample constitutes a copyright infringement requires knowledge of copyright law, making judgments about how much lexical similarity amounts to infringement, and the intended application of the data (Henderson et al., 2023a; Balganesh, 2012). Other approaches could resemble detection methods for data contamination, such as fuzzy string matching, or audio and video fingerprinting, as used by Youtube to identify copyrighted pieces (Cano et al., 2005; Wu et al., 2017).

Identifying problematic data without access to the training data. Regulators, auditors, or other entities who don't necessarily have access to a system's training data may need to find proxies for problematic data based on model behavior. Potential approaches include calculating confidence scores for the inclusion of data points (Li et al., 2024a), or using data watermarks introduced by creators that can be detected without access to the training corpora (Wei et al., 2024). Other approaches may be used to identify the use of copyrighted data, including inference attacks (Shokri et al., 2017) and influence functions (Grosse et al.,

2023; Choe et al., 2024). Yet, these approaches lack robustness – an issue that further research could aim to address.

Tracing data provenance. It can be challenging for model developers to ensure training data is correctly licensed due to licenses frequently being aggregated and misrepresented (Longpre et al., 2023). Hence, better tooling for data provenance will be necessary if developers are to feel comfortable that they are honoring creators' licenses. The Data Provenance Initiative (Longpre et al., 2023) conducted a large-scale audit of open-source fine-tuning data collections and cataloged corresponding data sources, licenses, and creators in an attempt to establish the provenance of data. However, the Data Provenance Initiative is limited in the datasets they cover, and expanding it to other data is resource-intensive. Automated license collection or standardized meta-data reporting for datasets (Longpre et al., 2024b) could help developers to release systems without facing legal ambiguity.

Silent removal of harmful data. Another challenge is that of being able to remove harmful data without the unwanted side-effect of publicly identifying it, and thus allowing malicious actors an easy way to source such data (Thiel, 2023). For example, the LAION-5B image dataset (Schuhmann et al., 2022) is composed of image URLs and metadata, as opposed to the images themselves. If one were to simply remove the URLs of the identified harmful images from the dataset, then this could provide malicious actors with the locations of these harmful images, either by comparing the datasets before and after removal, or directly through a repository change log where the dataset is hosted. However, care should be taken as methods for addressing this challenge could potentially also allow for the subversion of existing techniques (such as open change logs) for facilitating transparency into developers' data handling practices.

3.1.2 Infrastructure and Metadata to Analyze Large Datasets

Motivation: In addition to improved methods for auditing and identifying problematic samples in datasets, methods and infrastructure are needed for implementing these methods at scale. Contemporary datasets commonly contain on the order of tens of terabytes of data,[9] introducing the challenge of having the computational resources to store and handle such large quantities of data. Addressing this challenge will be crucial for enabling the identification of problematic data in practice.

Open problems:

Automating meta-data collection for datasets. A challenge with large-scale dataset infrastructure is that metadata – including links to the original data sources or license information – is not always provided (Piktus et al., 2023). An open research question is how the collection of metadata for previously published datasets could be automated. Additionally, exploring the automatic addition of cryptographic check sums as part of the metadata at the dataset level could help to enable users to confirm that the data they download matches the original data. Such an approach could be a partial solution to ensuring that datasets have not been altered (either maliciously or incidentally), especially in light of advances in data poisoning attacks (Carlini et al., 2023a).

Determining relevant targets for dataset analysis. Another question concerns the appropriate metrics for ascertaining the suitability of datasets for use in training, based on diffuse and macro-scale properties. For example, a dataset may be biased not as a result of the inclusion of individual samples, but the overall distribution of all included samples. Determining which measures and metrics are relevant in large-scale dataset analyses is an open challenge (Cho & Lee, 2021). A further open question concerns the information necessary for applying these metrics when evaluating datasets, extending the work on model and data cards by Mitchell et al. (2022b) and Pushkarna et al. (2022).

Developing search tools for large datasets. While there exists some work to quantitatively analyze dataset attributes (Mitchell et al., 2022b), this methodology generally "does not adapt well to web-scale corpora" (Piktus et al., 2023). One initial attempt to close this gap and to provide infrastructure for the quantitative and qualitative analysis of large datasets is the ROOTS Search Tool (Piktus et al., 2023). However, at the moment, ROOTS is limited to the 1.6TB corpus used to train BLOOM, and does not support other large-scale datasets. Extending this tool or creating similar tools for other open-access datasets could

[9] See for example, FineWeb (Penedo et al., 2024), a dataset of English text totaling 45 terabytes.

help with large-scale data governance. A related effort by Elazar et al. (2024) has provided an example of such an extension.

3.1.3 Attribution of Model Behavior to Data

Motivation: As noted above, one way in which training data may be problematic is if it causes undesirable downstream effects in models trained on that data, such as reproducing false information. In order to identify samples that should be excluded from training datasets for this reason, it may be necessary to understand how dataset composition can affect model performance. Thus, a complementary open issue to identifying problematic data is attributing model behavior to specific data points. Insights could be applied to different datasets used during development. For example, while pretraining data forms the basis for overall model performance (see e.g. Blakeney et al., 2024), preference data used during fine-tuning has a larger impact on the extent to which the resulting model is aligned with users' interests.

Open problems:

Understanding how pretraining data affects model behavior. Due to the size and complexity of contemporary AI systems, collective understanding of how specific training samples can contribute to problematic model behavior is incomplete (Lin et al., 2022a; Siddiqui et al., 2022). Udandarao et al. (2024) have, for example, investigated how pretraining concept frequency impacts downstream performance. Others have studied how upsampling domain-specific data relative to generic web-scraped text affects model performance (Blakeney et al., 2024).

Understanding properties and effect of preference data on fine-tuned models. In addition to pretraining data, future work could aim to understand the effects of the preference data used to fine-tune models, whether this is via a reward model, constitution, or another representation of preferences. It may be relevant to ensure that preference data has certain properties, such as representativeness, diversity, or neutrality, and design evaluations for the quality of this data. Related work in this context is research on preference aggregation for fine-tuning language models (Mostafazadeh Davani et al., 2022; Siththaranjan et al., 2024; Barnett et al., 2023) and scaling laws for reward model over-optimization (for example, Gao et al., 2023a).

Understanding the impact of synthetic data. Model developers may face data scarcity as current AI systems require increasing amounts of training data (Villalobos et al., 2022). Synthetic data generation offers an alternative to creating new authentic data, which can be both time-consuming and resource intensive. However, the impacts of training on synthetic data on model behavior are not yet well understood (Guo et al., 2023; Alemohammad et al., 2023; Martínez et al., 2023; Shumailov et al., 2023; Gerstgrasser et al., 2024). Questions remain on the effect that different strategies of training on synthetic data can have on model performance and bias, given that synthetic data has been shown to lack representativeness and insufficiently reflect imperfections of real-world data (Hao et al., 2024). Future research could aim to provide greater clarity regarding the effect of training on synthetic data, or develop uses of synthetic data for promoting model safety and reducing bias.

Balancing tractability and accuracy in data attribution. One key challenge in data attribution is the trade-off between computational tractability and accuracy (see, for example, Ghorbani & Zou, 2019; Jia et al., 2019; Ilyas et al., 2022; Akyürek et al., 2022), such as when using influence functions to attribute behaviors to data examples (Basu et al., 2020; Grosse et al., 2023; Choe et al., 2024). Park et al. (2023b) aim to overcome this trade-off by introducing TRAK (*Tracing with the Randomly-projected After Kernel*) but so far this work has only been tested for small foundation models. Due to TRAK's methodology requiring the training of multiple model versions for different subsets of the training set, it is unlikely that such a method would scale to the largest models, creating an avenue for future work.

3.2 Compute

Example Research Questions

13. What hardware properties or chip specifications are most indicative of suitability for AI training and/or inference? How does this differ from other scientific, business, or casual uses of high-end hardware? (3.2.1)
14. How efficiently can AI models be trained using a large number of small compute clusters? (3.2.1)
15. How can decentralized training attempts be identified? (3.2.1)
16. Can large training runs be detected while retaining developer privacy, for example through identifying signatures in processor utilization? (3.2.2)
17. Can compute workloads be reliably classified as either training, inference, or non-AI-related, for example through identifying signatures in processor utilization? (3.2.2)

3.2.1 Definition of Chip and Cluster Specifications for Model Training

Motivation: Compute governance has been proposed as a lever for governing advanced AI systems, due to the large amounts of computing resources required for their training and deployment (Sastry et al., 2024). However, despite its potential efficacy, compute governance is a blunt instrument, with previous actions including the restriction of the sale of a broad range of high-end chips to China (Bureau of Industry and Security, 2022b; 2024). It would thus be beneficial to be able to limit compute governance interventions to only the chips or compute clusters that are most relevant for developing and deploying AI systems of interest to policymakers. Ensuring that compute governance is targeted only where needed will require thoughtful derivation of metrics and specifications that capture hardware of concern, while excluding the vast majority of computing resources that are not used for industrial-scale AI.

Open Problems:

Assessing the effect of chip specifications on AI workload suitability. One open issue is understanding how different chip specifications – such as throughput, memory bandwidth, memory capacity, and interconnect bandwidth – affect a chip's suitability for different AI workloads, or the suitability of a cluster made up of those chips. Previous regulations concerning chips have arguably contained loopholes due to limitations in their technical specifications. For example, NVIDIA's A800, "while compliant with October 7 controls, is still capable of performing complex AI tasks, and was a top seller in China before the updated October 17, 2023 controls." (Reinsch et al., 2023)

Understanding the implications of decentralized training and cluster size. A related sub problem is understanding decentralized training, specifically the question of how efficiently AI models can be trained using multiple geographically disparate compute clusters. There is substantial technical work on developing decentralized training methods (see, for example, Douillard et al., 2024). Another open problem is the efficiency and cost impact of using a larger number of less powerful chips within a cluster, as opposed to using a smaller number of more powerful chips totaling the same theoretical throughput, sometimes known as *slicing*.

3.2.2 Classification of Workloads

Motivation: In addition to classifying hardware, it is also useful to classify computational workloads in order to identify potentially concerning or anomalous workloads. For example, Executive Order 14110 requires the reporting of training runs above a particular compute threshold (The White House, 2023a). Being able to classify workloads while preserving customer privacy could assist with such reporting, as well as a range of other governance goals, such as identifying compute usage trends, and audit trails for the development of powerful models (Heim et al., 2024; Shavit, 2023).

Open Problems:

Privacy-preserving workload classification. Compute providers, such as data center operators and cloud computing firms, typically already collect a wide variety of high-level data on customers and workloads (Heim et al., 2024). An open question is thus whether it is possible to use this data to develop reliable workload classification techniques, for example, determining whether a training workload exceeds certain compute thresholds, or whether an inference workload involves malicious cyberactivity (Commerce Department, 2024). Such techniques would need to account for changes in the hardware, software packages, and specific algorithms used in AI workloads over time.

Ensuring workload classification techniques are robust to adversarial gaming. Adversarial compute customers may try to obfuscate their activities to avoid workload classification, by introducing noise in the way computational resources are used, or breaking up workloads across multiple cloud accounts, providers, or computing clusters. Designing workload classification approaches that are robust to this kind of gaming, or otherwise are able to detect when this kind of gaming is occurring, is an open challenge (Egan & Heim, 2023; Heim et al., 2024).

3.3 Models and Algorithms

Example Research Questions

18. How can the thoroughness of evaluations be measured? (3.3.1)
19. How can potential blind spots of evaluations be identified? (3.3.1)
20. How can potential data contamination be accounted for when conducting evaluations? (3.3.1)
21. How can mechanistic analysis of model internals, such as weights, activations and loss landscapes on particular data, be used to improve understanding of a model's capabilities, limitations and weaknesses? (3.3.1)
22. How generalizable are mechanistic analyses across models? (3.3.1)
23. How can methods for red-teaming models be scaled and/or automated? (3.3.2)
24. How can the capabilities and risks of AI agents be evaluated? (3.3.3)
25. How can the capabilities and risks of networks of multiple interacting AI agents be evaluated? (3.3.3)

3.3.1 Reliable Evaluations

Motivation: A great deal of research is focused on evaluating AI models to measure their performance and identify capabilities and failure modes, including by state actors (Department for Science, Innovation & Technology & AI Safety Institute, 2024; UK AI Safety Institute, 2024; Anthropic, 2024b). Yet, state-of-the-art AI systems can exhibit unpredicted downstream capabilities that often evade evaluations (Shayegani et al., 2023; Carlini et al., 2023b; Schaeffer et al., 2024). The rapid advancement and widespread deployment of AI systems has prompted major regions, including the European Union, China, and the United States, to put forward requirements for evaluating, reporting, and mitigating risks associated with these systems (Reuel et al., 2024a). For instance, Article 55 of the EU AI Act mandates that providers of general-purpose AI models with systemic risk perform model evaluations using standardized protocols and tools, including adversarial testing, to identify and mitigate systemic risks (Council of the European Union, 2024). Despite jurisdictions mandating capability evaluations across various risk areas, there is a lack of technical clarity on how to perform these assessments comprehensively and reliably (Chang et al., 2024; Zhou et al., 2023a), and for some risks such evaluations simply do not yet exist (Weidinger et al., 2023). In addition, evaluations for decision-making systems in high-stakes settings will likely demand a higher level of confidence than other applications, but it is unclear how to determine the required level of rigor based on use case.

Open Problems:

Ensuring sufficient testing. Determining whether an evaluation procedure has identified all, if not most, of the vulnerabilities of a system is an open problem. This is especially relevant if the evaluation pertains to capabilities, such as deception and long-horizon planning, that could enable harmful forms of misuse or make systems hard to oversee or control (Park et al., 2024; Shevlane et al., 2023; Hendrycks et al., 2023; Kinniment et al., 2023; Li et al., 2024b; Phuong et al., 2024; Bengio et al., 2024). This issue is most acute in the case behavioral evaluations that directly test models' performance on benchmarks or test cases (Liang et al., 2023; Srivastava et al., 2023; Gao et al., 2023b; Wang et al., 2023a; Lee et al., 2023b; Biderman et al., 2024), as such evaluations can fail to fully inform evaluators about a model's capabilities, with particular uncertainty around capabilities that the model may lack (Casper et al., 2024a). Specifically, if a model does not exhibit a particular behavior during testing, it is challenging to determine whether this is due to the model genuinely lacking the underlying ability or if the evaluation method applied was insufficient to surface it (Wei et al., 2022; Zhu et al., 2023a), for example, due to prompt sensitivity (Zhu et al., 2023a; Sclar et al., 2023). This ambiguity can lead to an incomplete understanding of a model's true capabilities and potential risks (Barrett et al., 2023; Schaeffer et al., 2023; Raji et al., 2021), and means that behavioral evaluations can only offer a lower bound on a model's potential to exhibit harmful behavior (Goel et al., 2024; Casper et al., 2024b). This motivates research into how to expand the scope of behavioral evaluations as well as how to estimate the robustness and extensiveness of existing evaluations (Chan, 2024).[10]

Improving evaluations using mechanistic analysis. The shortcomings of behavioral evaluation techniques motivate additional work on evaluation methods that do not suffer from the same limitations. One proposed approach toward this is to study their internal mechanisms (Burns et al., 2022; Olah, 2023; Carranza et al., 2023; Casper et al., 2024a; Bereska & Gavves, 2024). Although evaluations that involve developing interpretations of a model's internal structure have seen a great deal of interest, they remain largely untested in practice (although see Templeton et al., 2024; Gao et al., 2024). Furthermore, interpreting the internal mechanisms of models inevitably involves reductions in the complexity of the system being studied (Gilpin et al., 2018).

Testing the validity of evaluations. It can be hard to have high confidence that the results of the evaluations reflect properties of the model rather than the evaluation methodology employed – that is, that the evaluation is internally valid.[11] Making progress on this could be challenging due to uncertainties pertaining to both evaluations and the models to which they are applied, making it difficult to attribute results to either the evaluation or the model. One potential avenue for progress could be using *model organisms* – smaller simpler AI models that have particular properties by construction (for example Hubinger et al., 2024) – to test whether evaluations for the constructed property are able to reliably detect it. However, this method would likely not be of use if trying to develop evaluations for model properties that are only exhibited by the most capable models. More generally, designing meta-evaluations that assess the reliability and consistency of evaluation methodologies across different models and contexts remains an open research challenge.

Establishing a causal relationship between procedural design choices and system characteristics. Similar to attributing a model's properties to characteristics of its training data (see Section 3.1.3), it may be possible to attribute behavior or performance to design decisions made during a model's development process (see, for example, Simson et al., 2024). Establishing causal relationships between such decisions and resulting system properties may allow for greater standardization of development best practices.

Understanding potential risks and capabilities of future AI systems. Finally, it is also difficult to study certain risk scenarios that might emerge through advanced capabilities – for example, the ability to plan over a long horizon – due to their hypothetical nature. It could be useful to develop and study specific demonstrations of harmful behavior and the effectiveness of current safety techniques against these behaviors in current AI systems, similar to the model organisms approach suggested in the previous paragraph (Scheurer et al., 2024; Järviniemi & Hubinger, 2024; Hubinger et al., 2024).

[10]Beyond insufficient model-level testing, evaluations along the whole AI life cycle are lacking. We address this problem separately in Section 3.4.

[11]Internal validity refers to the extent to which an evaluation accurately measures what it intends to measure within its specific context, while external validity refers to how well the evaluation results can be generalized or applied to other situations, populations, or contexts beyond the original study.

3.3.2 Efficient Evaluations

Motivation: Ideally, it would be possible to test an AI system under all possible inputs to ensure that it would not produce a harmful output for any of them. However, performing a brute-force search over possible inputs is intractable due to the astronomically large input spaces for modern systems,[12] and the fact that whether an output is harmful may be unclear or context-dependent. As a result, current evaluation methods manually apply heuristics to guide vulnerability searches towards regions of the input space assumed to be more concerning – as observed in voluntary audits by developers (OpenAI et al., 2024; Kinniment et al., 2023; Touvron et al., 2023; Anthropic, 2023b). However, manual attacks quickly become impractical, expensive, and insufficient for conducting scalable evaluations (Ganguli et al., 2022), especially when searching across modalities and languages (Üstün et al., 2024). This problem is exacerbated for increasingly capable, general-purpose systems that have a significantly larger attack surface than narrower systems. The development of more efficient automated approaches for identifying model vulnerabilities will be crucial if results from model evaluations are to be applied as inputs to governance-relevant decisions.

Open Problems:

Making comprehensive red-teaming less resource-intensive. Recent progress has been made on automated red-teaming by using generative AI models to produce test cases (Perez et al., 2023; Shah et al., 2023), develop adversarial prompts (Deng et al., 2022; Perez et al., 2022; Mehrabi et al., 2023; Hubinger et al., 2024; Casper et al., 2023; Hong et al., 2024), and automatically evaluate the outputs of other models (Zheng et al., 2023; Chiang et al., 2023; Ye et al., 2023; Kim et al., 2023; Chao et al., 2024; Souly et al., 2024). Furthermore, automated search methods have been applied to find adversarial attacks as a way of enhancing or replacing manual methods (Wallace et al., 2019; Song et al., 2020; Shin et al., 2020; Guo et al., 2021; Shi et al., 2022; Kumar et al., 2022; Wen et al., 2023; Jones et al., 2023; Zou et al., 2023b; Liu et al., 2023b; Zhu et al., 2023b; Andriushchenko et al., 2024). However, despite these approaches for automating evaluations, thorough red-teaming remains a labor-intensive process, and many failure modes still evade red-teaming efforts (Shayegani et al., 2023; Carlini et al., 2023b; Longpre et al., 2024a). These challenges in part stem from how existing automated techniques are often computationally expensive and crude, requiring a large degree of guidance from human engineers (Mazeika et al., 2024). Qualitatively different approaches that automate some or all the red-teaming process – perhaps through the use of agentic AI systems that can plan, use tools, and dynamically evaluate systems – could also allow for more scalable evaluation.

3.3.3 (Multi-)Agent Evaluations

Motivation: *Agentic* AI systems are generally characterized by an ability to accomplish tasks from high-level specifications, directly influence the world, take goal-directed actions, and perform long-term planning (Chan et al., 2023b; Durante et al., 2024; Huang et al., 2024a). These capabilities could allow agentic systems to perform tasks with little human involvement and control. While economically useful, for example, as customized, personal assistants, or for autonomously managing complex supply chains, agentic systems could pose unique risks due to their ability to directly act in the world, potentially with difficult-to-predict impacts (Chan et al., 2023b; Lazar, 2024; Gabriel et al., 2024; Bengio et al., 2024).

Open Problems:

Evaluating and monitoring agentic systems. User customizability, such as through prompting or the integration of new tools, makes it particularly difficult to foresee the use cases and potential risks of agents (Shavit et al., 2023; Kolt, 2024; Cohen et al., 2024a), motivating potential measures for tracking and monitoring their actions (Chan et al., 2024). Furthermore, evaluating agents is a nascent field with significant challenges – existing agent benchmarks often don't have adequate holdout datasets, causing existing agents to game and overfit to the benchmark, which in turn results in unreliable evaluations of these systems (Kapoor et al., 2024b). Similar to non-agent benchmarks (see above), best practices are currently lacking, leading to inconsistencies across evaluations and limiting their reproducibility (Kapoor et al., 2024b). Thus, future work could aim to introduce best practices for evaluating agentic systems.

[12] For example, there are vastly more possible 20-token strings of text or 10×10 pixel images than there are particles in the observable universe. (One current estimate for the number of particles in the observable universe is 10^{80}. GPT-2's tokenizer had a vocabulary size of $50,257$, meaning that there are approximately $50,000^{20} \approx 10^{94}$ unique strings of 20 tokens. There are $256^{100} \approx 10^{240}$ possible 10×10 gray-scale images with integer pixel values in the range $[0, 255]$.)

Expanding limited multi-agent evaluations. On top of the difficulties of studying single-agent systems, multi-agent interactions add an additional layer of complexity due to information asymmetries, destabilizing dynamics, and difficulties in forming trust and establishing security. These problems can lead to unique complexity and failure modes (Hammond et al., 2025; Chan et al., 2023a; Akata et al., 2023; Mukobi et al., 2023). In addition, it may be the case that collectives of agents exhibit unpredictable capabilities or goals not attributable to any one agent in isolation (Hammond et al., 2025). If AI agents become increasingly embedded in real-world services, such as in finance or the use of web services, it will be relevant to understand such multi-agent dynamics.

Attributing downstream impact to individual agents. For issues of liability, it will be critical to be able to determine which agent(s) or system(s) can be held responsible for a particular decision or action, if any. This may be complicated for cases in which the cause is not solely attributable to a single AI agent.[13] Having methods of tracing multi-agent interactions and determining the cause of a particular outcome could help to solve this problem. An open technical question in this context regards techniques for monitoring individual agents' contributions to multi-agent systems, in order to ease attribution of responsibility (Friedenberg & Halpern, 2019).

3.4 Deployment

Example Research Questions
26. How can the downstream societal impacts of AI systems be predicted and/or determined? (3.4.1)
27. How can downstream impact evaluations be scaled across languages and modalities? (3.4.1)
28. How can benchmarks be designed in a way that ensures construct validity and/or ecological validity? (3.4.1)
29. How can dynamic simulation environments be designed to better reflect real-world environments? (3.4.1)

3.4.1 Downstream Impact Evaluations

Motivation: The performance of models in isolation is an imperfect proxy for the impact that AI systems will have in everyday use. Thus, comprehensively understanding the impacts that AI could have on society demands robust methods for evaluating systems in dynamic, real-world settings (Ibrahim et al., 2024). Having such methods would allow policymakers a higher-fidelity picture of where governance intervention might be necessary in order to address potential harms.

Open Problems:

Predicting the downstream societal impacts of AI systems. Although understanding overall societal impact is an overarching goal of much work on evaluating AI, it is a difficult, large-scale sociotechnical problem (Dolata et al., 2022; Solaiman et al., 2023; Rakova & Dobbe, 2023; Weidinger et al., 2023; Dobbe & Wolters, 2024; Bengio et al., 2024). For example, simplified technical proxies for complex concepts like *fairness* and *equity* insufficiently measure the disparate impact of AI systems on diverse communities (Blodgett et al., 2020; Selbst, 2021). It may also be logistically challenging to conduct downstream evaluations. For example, while it is known that large language models exhibit significant cross-lingual differences in safety and capabilities (Yong et al., 2023; Wang et al., 2023c;b; Jin et al., 2023; Üstün et al., 2024), it is expensive and time-consuming to thoroughly evaluate the cross-lingual properties of models due to the required coordination between speakers of many languages. Ultimately, thoroughly assessing downstream societal impacts requires nuanced analysis, interdisciplinarity, and inclusion (Hagerty & Rubinov, 2019; Bengio et al., 2024). While recent work has taxonomized societal impacts and provided an overview of early techniques for their

[13] See, for example, (Wex Definitions Team, 2023)

evaluation (Moss et al., 2021; Shelby et al., 2022; Raghavan, 2023; Solaiman et al., 2023; Weidinger et al., 2023; 2024), there remains a lack of structured, effective methods to quantify and analyze these impacts.

Ensuring construct validity of evaluations. It can be difficult to establish confidence that the proxy used in an evaluation or benchmark accurately captures the concept it aims to measure – that is, its construct validity (Raji et al., 2021; Bowman & Dahl, 2021; Hutchinson et al., 2022; Subramonian et al., 2023; McIntosh et al., 2024). For example, while MMLU (Hendrycks et al., 2021) claims to assess a model's understanding and memorization of knowledge from pretraining through the proxy of performance on question answering, it is unclear how well the ability to accurately answer questions serves as an indicator for understanding. Future research could aim to evaluate the construct validity of current benchmarks as well as ensure construct validity for AI system evaluations, perhaps taking inspiration from prior work in psychology (see, for example, (Westen & Rosenthal, 2003; Strauss & Smith, 2009; Smith, 2005)).

Ensuring ecological validity of evaluations. In addition to construct validity, establishing the ecological validity of benchmarks is an open challenge.[14] Current benchmarks tend to be biased towards easily-quantifiable and model-only metrics, potentially making them ill-suited for predicting how well models perform when deployed in real-world settings (Ouyang et al., 2023; Lee et al., 2023a). Future work could aim to assess the correlation between performance on benchmarks and downstream performance, as well as propose evaluation methods for which this correlation is tighter. Another consideration is that benchmarks used for assessing capabilities are oftentimes used to guide development of models, limiting the extent to which such benchmarks can be seen as unbiased (Marie et al., 2021; Dehghani et al., 2021; Madaan et al., 2024; Salaudeen & Hardt, 2024), and making it challenging to know how to interpret resulting scores.

Designing dynamic evaluations and real-world simulation environments. Despite many user interactions with AI systems taking place in dynamic, multi-turn environments, current benchmarks only evaluate performance in such settings to a limited degree (Ibrahim et al., 2024). By creating dynamic evaluation frameworks (for example, Dynabench, 2023; Park et al., 2023a), researchers could better assess a system's performance, inherent characteristics, and potential risks in a more realistic manner compared to static test sets. This would require significant investment in infrastructure and tooling, such as sophisticated simulated environments tailored to specific domains like hacking, persuasion, or biosecurity. Similarly, experiments with human subjects are valuable to understand the risks from human capability increases through AI models (Mouton et al., 2023), but are less common due to their resource-intensiveness.

[14] *Ecological validity* refers to the extent to which results from experiments generalize to contexts outside of the testing environment. Compare this to *construct validity* which refers to the extent to which an assessment measures the target construct.

4 Access

Many governance actions will likely require third-party access to system components, along with the provision of resources such as compute. As examples, external access will likely be a necessary consideration for facilitating third-party audits (Raji et al., 2022b; Anderljung et al., 2023b; U.S. National Telecommunications and Information Administration, 2024; Mökander et al., 2023; Casper et al., 2024a), evaluations by government AI safety institutes (GOV.UK, 2023; National Institute of Standards and Technology (NIST), 2024; Department for Science, Innovation and Technology et al., 2023b), and independent academic research (Bucknall & Trager, 2023; Kapoor et al., 2024a; Longpre et al., 2024a; House of Commons Science, Innovation and Technology Committee, 2024).

There are numerous reasons for why it may be desirable to enable these functions to be performed by those other than the system developers. Firstly, actions such as evaluation and auditing could benefit from the independence of being conducted by parties external to AI developers, so that developers do not "mark [grade] their own homework" (Gerken & Rahman-Jones, 2023). Furthermore, AI developers may not have the capacity or incentives to conduct research to the extent needed for advancing our scientific understanding of AI systems at a rate comparable to that of advances in AI development, motivating the need for involvement of the broader academic community. Third-party access may also allow for broader cultural diversity and representation in the development and governance of AI (Dobbe et al., 2020; Delgado et al., 2023; Held et al., 2023; Crowell, 2023), though should not be seen as a sufficient measure (Chan et al., 2021; Sloane et al., 2022).

However, many of the aforementioned actions are precluded due to insufficient external access to relevant system components, especially in the case of the most state-of-the-art systems (Solaiman, 2023; Bommasani et al., 2023a; Bucknall & Trager, 2023; Casper et al., 2024a). This is often due to concerns regarding developers' intellectual property, privacy of data subjects, legal uncertainty, and the safety of the system in question (Seger et al., 2023).

Access to compute and other resources peripheral to the systems under consideration, such as training data, is also essential for many auditing and research functions (Ahmed & Wahed, 2020; Besiroglu et al., 2024; Ojewale et al., 2024). The challenges associated with facilitating access to such resources will also need to be addressed if these functions are to be fulfilled in academia and the public sector.

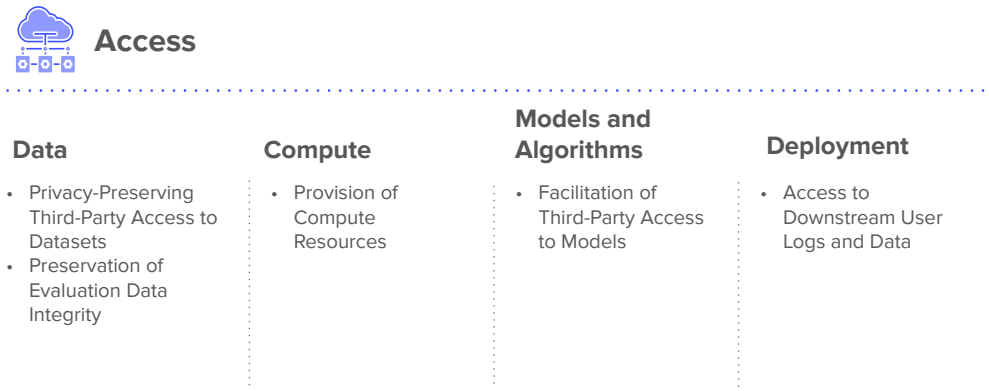


Figure 3: Open problem areas in the *Access* capacity, organized by target

4.1 Data

Example Research Questions
30. How can data access be structured so as to preserve privacy while enabling meaningful auditing? (4.1.1)
31. How can data access be reconciled with privacy-preserving machine learning? (4.1.1)
32. How can openly hosted datasets be prevented from contaminating training data? (4.1.2)
33. How can independent evaluation on standardized datasets be facilitated without openly hosting evaluation datasets? (4.1.2)

4.1.1 Privacy-Preserving Third-Party Access to Datasets

Motivation: Access to training datasets is crucial for enabling external data audits that aim to identify instances of harmful, personal, or inappropriate data being included in datasets (Thiel, 2023; Birhane et al., 2021; 2023; Subramani et al., 2023; Luccioni & Viviano, 2021, see also Sections 3.1 and 5.1). External access to datasets could also be instrumentally necessary for facilitating assessment and research into models (Bucknall & Trager, 2023; Ojewale et al., 2024), for example, as research into how the content of training datasets influence downstream model behavior depends on visibility into the training datasets (for example Udandarao et al., 2024, see also Section 3.1).

However, naïvely providing unrestricted access to datasets may violate legal bounds, for example by reproducing copyrighted data, or otherwise raise security concerns. Additional challenges stem from the lack of legal clarity regarding how to define problematic data, such as what constitutes a copyright violation in the context of AI training (Henderson et al., 2023a; Quang, 2021). Furthermore, developers may be reluctant to provide unrestricted access to proprietary datasets due to the high costs associated with collating such datasets,[15] the risk of leaking sensitive intellectual property, or, in some cases, risking disclosing having knowingly trained on illegally collected data. These concerns may be able to be alleviated by providing greater transparency into the makeup, contents, and aggregate characteristics of datasets, as well as information about sources and compilation methodologies, in the absence of unrestricted access to the entire dataset.

Open problems:

Structuring external access to datasets to reduce privacy risks. Research could aim to develop methods for providing sufficiently deep access to datasets, for example, for the purpose of auditing and evaluation, while protecting privacy of data subjects. Existing work in this direction includes Trask et al. (2023) which suggests a framework for allowing third-parties to propose and execute approved queries on AI systems and third-party data without exposing sensitive information beyond the explicitly authorized results. Similarly, *Project Oak* is a software package for providing "security and [...] transparency around how data is used and by whom, even in a highly distributed system", through the use of trusted execution environments (TEEs) on specialized hardware (Project Oak, see also Section 6.2). Inspiration can also be taken from other industries, including healthcare (NHS Research SDE Network, 2024). For example, *OpenSAFELY* provides a platform for the analysis of patient healthcare data for the purposes of academic research through a combination of pseudonymization, methods for working with data *in situ* while obfuscating raw patient data, and providing transparency into researchers' use of the platform (Bennett Institute for Applied Data Science, 2024). Future research could draw insight from this case to propose methods for safely accessing user data for research into the societal impacts of AI.

Addressing the tension between data access and privacy-preserving machine learning. A significant body of research has explored how we can train machine learning models when data-owners do not trust

[15] Including the cost of data collection and the resources to ensure that licenses are correctly represented to avoid incorrect infringements.

model-trainers – for example through federated learning (Kairouz et al., 2019) and training on encrypted data (Xie et al., 2014; Nandakumar et al., 2019). In such cases, it can be challenging to provide data access given that the developer themselves lacks such visibility. Future work could explicitly address this tension, proposing methods that allow for the auditing of data that may have been encrypted during training.

4.1.2 Preservation of Evaluation Data Integrity

Motivation: Current standardized methods for evaluating models often utilize openly-available datasets, with the goal of comparing model performance like-for-like (Reuel et al., *forthcoming*). At present, this is largely achieved by hosting such datasets openly in online repositories such as *HuggingFace*[16] (see, for example, Hendrycks et al., 2021; Srivastava et al., 2023; Gao et al., 2023b). However, openly hosting evaluation datasets online risks their inclusion in web-scraped training datasets (Deng et al., 2023), either accidentally, or intentionally as a method for artificially inflating benchmarking results. Such contamination of training data has serious implications for the efficacy and reliability of these standardized metrics (Oren et al., 2023; Roberts et al., 2023; Jiang et al., 2024; Zhang et al., 2024; Schaeffer, 2023)

Open problems:

Identifying and mitigating contamination of training datasets. Current approaches to mitigating contamination of training datasets are rudimentary. A potential post-contamination approach is to detect data contamination (Dong et al., 2024b; Golchin & Surdeanu, 2023) and then correct for it when scoring benchmarks. For example, OpenAI and Meta measured which benchmarks' test samples were potentially included in the pretraining data of GPT-4 and Llama 2, respectively, and reported how scores differed between contaminated test samples and non-contaminated test samples (OpenAI et al., 2024; Touvron et al., 2023). Zhang et al. (2024) attempted to correct for data contamination by creating a version of the benchmark GSM8k that is comparable in terms of tasks, complexity and human solve rates, and reported which models' scores dropped precipitously. Other approaches pertain to the design of benchmarks that are robust against contaminated models by using templates from which variations of a task can be generated (Yu et al., 2023; Srivastava et al., 2024) or by being frequently updated (White et al., 2024). However, the frequent updating approach is resource intensive, especially if covering a large span of test tasks and fields, and designing templates to support variations of tasks may not be feasible for tasks that don't follow a predictable structure. Alternatively, (Srivastava et al., 2023) make use of a *canary string*, that is, a globally unique identifier that is included in all sub repositories of the BIG-bench collection in order to ease identification of these test samples in training datasets. BIG-bench also includes a dedicated `training_on_test_set` task, which serves as a "post-hoc diagnosis of whether BIG-bench data was used in model training" (Srivastava et al., 2023). However, use of canary strings depends on any and all copies of the repositories to also include the string, and thus is not robust to negligent users.

Evaluating on private or encrypted evaluation datasets. Alternatively, research could aim to develop ways in which models can be independently evaluated on a private or encrypted test set. Indeed, some recent benchmarking datasets have only been made available only through a custom evaluation API (see, for example, Sawada et al., 2023). Furthermore, popular dataset repositories including *HuggingFace*, as well as competition platforms such as *Kaggle*, gate access to evaluation datasets in order to reduce the risk of contaminating training data.[17] Finally, Bricman (2023) proposes *hashmarks*, a "protocol for evaluating language models in the open without having to disclose the correct answers" by cryptographically hashing a benchmark's reference solutions before publication. Further work could develop this, and similar protocols for reliably evaluating system capabilities on private evaluation data.

[16] https://huggingface.co/
[17] (Hugging Face, 2024; kaggle, 2024)

4.2 Compute

Example Research Questions
34. How can public compute resources be allocated fairly and equitably between users? (4.2.1)
35. How can public compute infrastructure be developed in a way that ensures interoperability between models and software packages? (4.2.1)
36. How can assurance be given that researcher compute provisions are being used for intended and stated purposes? (4.2.1)

4.2.1 Addressing Compute Inequities

Motivation: Compute usage by private companies in training and running models has increased exponentially in the past years, and now greatly exceeds the compute resources available for non-industry researchers (Maslej et al., 2024; Besiroglu et al., 2024). While some researchers have found that the majority of academic researchers do not feel primarily constrained by compute access (Musser et al., 2023), others have found that this access inequality is specifically limiting researchers' contribution to frontier research (Ahmed & Wahed, 2020; Besiroglu et al., 2024; Birhane et al., 2023). To address these concerns, there have been proposals and funding for public compute infrastructure (NDIF proposal; Ho et al., 2021; Organisation for Economic Co-Operation and Development, 2023; National Artificial Intelligence Research Resource Task Force, 2023; UK Research and Innovation, 2023). Though the success of these initiatives primarily depends on raising funds to purchase sufficient compute, technical advances could still be instrumental in their success.

Open problems:

Ensuring interoperability of public compute resources. Public compute resources should be compatible and interoperable with a wide range of models and software packages, in order to support the range of research projects that would be conducted. System performance can vary considerably depending on the hardware and software on which it is run (Nelaturu et al., 2023; Gundersen et al., 2022), and common ML software frameworks can lose more than 40% of their key functionality when ported to non-native hardware (Mince et al., 2023). Future research could aim to propose solutions that address these observed defects.

Ensuring environmental sustainability of public compute resources. The resourcing requirements of large-scale supercomputers and data centers are considerable, both in terms of energy (Strubell et al., 2019) and other resources such as water, used for cooling (Mytton, 2021). Thus, measures will need to be taken to balance broad access to computing resources with environmental sustainability. The environmental impacts of AI systems and associated open problems is discussed further in Section 8.

Ensuring public compute is used for intended purposes. System administrators would need to be able to ensure that public compute resources are being used for the stated purposes, rather than malicious or otherwise unintended uses, for example by performing "workload classification" (Heim et al., 2024, see Section 3.2 for open problems in this context). Such oversight methods would need to preserve end-users' privacy to system administrators, as well as that of any potential subjects of data used in conducting experiments (see Section 5.2).

Equitably allocating public compute resources. Given the high demand for public compute resources, another issue will be the efficient and fair allocation of processor time between users. While methods for allocating compute resources in more general cases have been explored (Ghodsi et al., 2011; Wang et al., 2015; Xu & Yu, 2014; Souravlas & Katsavounis, 2019; Jebalia et al., 2018), future work could aim to ensure their applicability to this case. Alternatively, research could aim to find AI-specific optimizations for allocating compute among diverse users.

4.3 Models and Algorithms

Example Research Questions
37. What research and auditing methodologies are possible given a range of forms of access on the continuum between black- and white-box access? (4.3.1)
38. How do different forms of access affect potential risks of misuse of models? (4.3.1)
39. How do different forms of access on the continuum between black- and white-box access affect the risk of model theft or duplication? (4.3.1)
40. How can model access requirements for research and auditing be reconciled with commercial and/or safety concerns? (4.3.1)

4.3.1 Facilitation of Third-Party Access to Models

Motivation: A fundamental requirement of conducting external research and evaluation of AI systems is having access to the underlying models. However, many systems are not released openly (Solaiman, 2023), and, while access requirements vary widely for different external actors, current APIs do not offer sufficient depth or flexibility of access to facilitate many actions important for research and evaluation (Bucknall & Trager, 2023; Casper et al., 2024a; Longpre et al., 2024a). For example, Casper et al. (2024a) argue that evaluations conducted with solely black-box access can "produce misleading results" and only offer "limited insights to help address failures" due to their not revealing complete information regarding the nature of discovered flaws. It is a challenge to find the balance between providing external parties with sufficient access for conducting independent research and evaluation, while addressing developers' concerns such as IP theft or misuse of their models. While there are certainly social and legal pathways that can be pursued towards this end, there are also several technical avenues (Casper et al., 2024a).

Open problems:

Illuminating the continuum between black- and white-box access. Greater clarity regarding how much, and what kinds of, research can be conducted with different depth and breadth of access would be helpful for navigating the trade off between access and security (Bucknall & Trager, 2023). Different auditing procedures can demand varied levels of access, motivating the need for a range of methods for supporting researchers, rather than prescribing a set approach (Casper et al., 2024a). On the flip side, a clearer picture of how differing forms of access bear on developers' security and privacy concerns would also be needed. Black-box access already allows for the training of *distilled models* that can then be used to generate effective adversarial attacks against production models via transfer (Zou et al., 2023b), and fine-tuning APIs can be ineffective at guarding against the removal of pre-deployment safety measures (Qi et al., 2023). Meanwhile, the ability to view language model output logits has been shown to be sufficient for extracting proprietary system information, including the model's hidden dimension, though it is unclear the extent to which this is a practical threat (Carlini et al., 2024). Further research could aim to elucidate how the provision of intermediate forms of grey-box access could exacerbate these existing vulnerabilities when compared to the baseline of black-box access.

Applying technical measures to address vulnerabilities of greater access provisions. Outstanding technical questions regarding external model access include whether the use of technical tools, such as privacy-enhancing technologies or TEEs, could enable near-white-box auditing and research while addressing commercial and safety concerns. For example, Aarne et al. (2024) describe how an approach combining multi-party computing with TEEs "could be used by a third-party evaluator to run tests on an AI model without ever having direct access to the unencrypted weights." Future research could assess the extent to which such solutions allow model providers, auditors, and regulators to interact in a way that is: easy to set up for the model provider; leaks no model information to the auditor; leaks no audit information to the model provider; and does not compromise the cybersecurity of the model provider. Alternatively, work could aim to incorporate approaches for providing third-party model access at scale into secure and trusted compute

clusters, including public compute resources (Anderljung et al., 2022; Heim, 2024). To date, there has been preliminary work on protocols for secure privileged evaluations (Trask et al., 2023), though there is a lack of existing applications or established best practices.

Ensuring version stability and backward-compatibility of hosted models. Large commercial models are frequently and continually updated during deployment, with prior versions often being replaced without notice or knowledge. However, reproducibility and replicability of independent research conducted on proprietary models depends on stable and continued access to models, even after their being succeeded by newer versions (Pozzobon et al., 2023; Bucknall & Trager, 2023; Biderman et al., 2024). Maintaining hardware and software compatibility (Mince et al., 2023) may be necessary in order to be able to provide access to discontinued systems upon request. Future work could lay out best practices for documenting and communicating when models are being deprecated or discontinued, in order to ease reproducibility concerns.[18]

4.4 Deployment

Example Research Questions
41. How can user logs and data be used for downstream impact assessments while preserving the privacy of data subjects? (4.4.1)
42. How could responsibilities for providing user data access be effectively allocated along the AI value chain? (4.4.1)
43. What cryptographic methods can be developed to allow analysis of user interaction data without revealing individual user identities or sensitive information? (4.4.1)
44. How could secure multi-party computation be leveraged to allow collaborative analysis of user logs across different entities in the AI value chain? (4.4.1)

4.4.1 Access to Downstream User Logs and Data

Motivation: Assessing models post-deployment, as covered in Section 3.4, requires access to relevant real-world data on user interactions with systems (Nicholas, 2024). This data can be used to directly assess aspects of user-model interactions, build evaluations that are more reflective of real-world use, and guide assessments of societal-level patterns in key sectors (Ibrahim et al., 2024). While there are crowd-sourcing initiatives which allow users to voluntarily submit some of their interaction data to create research datasets (The Allen Institute for Artificial Intelligence, 2024; ShareGPT, 2022), to the best of our knowledge, no model provider has made their interaction datasets, or privacy-preserving metadata about these logs, widely available.

Access to real-world user data, such as usage logs and user feedback records, may be additionally relevant for legal purposes. For example, legal cases may arise when users experience harm as a result of an interaction with an AI system – either directly as a participant in the interaction of concern, or indirectly as a subject of actions taken following an interaction. In such cases, the availability of information including user logs and audit trails to prosecutors or courts may be relevant for determining the outcome of a case.

Open problems:

Addressing user privacy concerns regarding access to user logs. External access to user interaction data must overcome privacy concerns relating to the collection, sharing, and analysis of potentially sensitive and identifying user information. This challenge has parallels in other industries, for example in online platform governance, where the EU's Digital Services Act (Digital Services Act) mandates independent researcher access to platform data (EU Joint Research Centre, 2023; Albert, 2022). While implementation challenges remain in the case of the Digital Services Act (Leerssen, 2021; Leerssen et al., 2023; Leerssen, 2023; Jaursch et al., 2024; Morten et al., 2024), more developed solutions can be seen in the healthcare

[18]See (Luccioni et al., 2022) for related work on deprecating datasets.

sector (see the discussion of OpenSAFELY in Section 4.1). Inspiration could be taken from these sectors for ensuring user privacy while providing access to user logs for research purposes.

Understanding how access responsibilities may vary along the AI value chain. Additional difficulties in providing user data stem from the complexities of the AI value chain (Küspert et al., 2023) – for example, in the case that a foundation model is built upon and incorporated in a user-facing application by a downstream deployer. In this scenario, it is not immediately clear how access requirements interact with the division of information between the foundation model developer and subsequent deployer. An additional challenge may emerge if the provision of user data in this case implicitly reveals information about how the deployer is using the model as part of their service, potentially putting their IP at risk of leakage. Further work could clarify potential access responsibilities in this and similar situations.

5 Verification

In many cases, it may be beneficial to be able to verify claims regarding AI systems' properties, capabilities, and safety as a way of increasing trust between actors (Brundage et al., 2020). While related to assessment, covered in Section 3, verification concerns the process of checking whether an AI system "complies with a [specific] regulation, requirement, specification, or imposed condition" (IEEE, 2011), as opposed to evaluating the system's performance, capabilities, or potential societal impacts. For example, an assessment task could be to uncover details about the data that a given model was trained on. In contrast, a verification problem could be, given a dataset and a model, to confirm or refute the claim that the model was trained on the dataset.

There is a trend towards an increasing amount of regulation and corresponding requirements for model developers, deployers, and users being passed by major national and international jurisdictions (Maslej et al., 2024). It may thus be necessary for model developers and deployers to verify and attest to certain properties of their AI systems in order to prove that they comply with regulations. On the flip side, governments may need to be able to verify whether actors in the AI ecosystem comply with the regulations, or verify that other countries are in compliance with international rules (Baker, 2023; Shavit, 2023; Avenhaus et al., 2006).

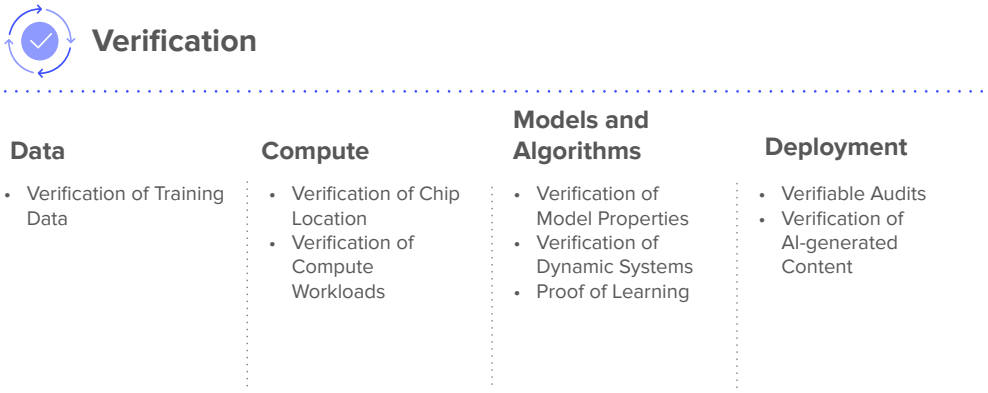


Figure 4: Open problem areas in the *Verification* capacity, organized by target

5.1 Data

Example Research Questions
45. How can it be verified that a model was (not) trained on a given dataset? (5.1.1)
46. How can it be verified that a dataset has certain properties, or (does not) include certain information? (5.1.1)
47. How can membership inference attacks be optimized for large-scale verification of training data in black-box settings? (5.1.1)
48. How could the verification process for correct use of licensed data in AI model training be formalized? (5.1.1)

5.1.1 Verification of Training Data

Motivation: Being able to verify the data on which a given model was trained – either as a model developer or third-party auditor – could aid in demonstrating compliance with data handling standards and regulation, including the Blueprint for an AI Bill of Right, or the EU's General Data Protection Regulation (The White

House Office of Science and Technology Policy, 2023; European Commission, 2016). In particular, even if it is possible to demonstrate that a dataset does not contain harmful or copyrighted material (see Section 3.1), this is insufficient for guaranteeing that a model was not trained on problematic data, as the developer may have used an alternate dataset to the one assessed. Being able to *post hoc* provide evidence that a model was trained on a specific dataset would help to preclude such instances.

Open Problems:

Verifying datasets used to train a model. Choi et al. (2023), building on Shavit (2023), formalize the *proof-of-training-data problem*, wherein a prover aims to "prove to a verifier that the resulting target model weights W^* are the result of training on data D^*," and propose a solution. However, the authors concede that their approach is not robust to all potential attacks, in particular additions of small amounts of harmful data, such as that used for inserting backdoors (Xu et al., 2021). In addition, Choi et al.'s protocol is not applicable to models for which data is not fully known before training, as in the case of online or reinforcement learning. Finally, their protocol requires that the "Prover disclose confidential information to the Verifier, including training data, model weights, and code" (Choi et al., 2023), which may create IP and security concerns. An alternate approach to verifying training data could be the application of *membership inference attacks* which aim to infer whether a given data point was contained in a given model's training data in a black-box setting (Shokri et al., 2017; Duan et al., 2024a; Wei et al., 2024). Future research could aim to suggest robust methods for verifying training data, assess the robustness of existing methods, or address the parallel issue of verifying that a given dataset (or sample) *was not* used in the training of a given model.

Verifying fair data use. Verifying copyright compliance and fair data use is a complex challenge that may require additional legal and technical frameworks (see Section 3.1) to extend or replace the Proof-of-Training-Data approach (Choi et al., 2023). Open challenges include formalizing the verification of correct use of licensed data, as well as verifying the exclusion of specific licensed data from training sets, should the respective license not allow the data to be used for training a model.

5.2 Compute

Example Research Questions

49. How can the location of AI hardware be verified? (5.2.1)
50. How can on-chip geolocation mechanisms be made robust to existing GPS spoofing methods? (5.2.1)
51. Can TEEs be used to robustly attest to the identity of the specific chip, or the data that it is processing? (5.2.2)
52. What methods can be used to verify compute usage without the use of TEEs? (5.2.2)
53. How can TEEs and their applications be designed in a way that limits their potential for misuse, for example through unnecessarily-broad surveillance? (5.2.2)
54. How can the computational overhead of verification mechanisms be reduced to a level that enables application across large compute clusters? (5.2.2)

5.2.1 Verification of Chip Location

Motivation: High-end data center AI chips are the subject of U.S. export controls, but are at present straightforward to smuggle (Huang, 2024; Harithas, 2024; Fist & Grunewald, 2023). One key technical problem for enforcement is that it's not currently possible to know the location or owner of a chip after it has been exported. Being able to verify a chip's location could also help users of cloud computing validate that their (or their customers) data is being processed in accordance with local data processing laws.

Open Problems:

Verifying chip location. Accurately detecting a chip's location remains an open challenge. One approach could be to measure verifiable latencies between the chip in question and a network of trusted servers (Aarne et al., 2024; Brass & Aarne, 2024). Other methods could be valuable for verifying that a large number of chips are co-located in a single data center, and have not been resold or dispersed. It may be possible to verify this using a proof of work challenge that would require a large number of co-located systems to complete in the allotted time (Jakobsson & Juels, 1999). Alternatively, chips could more directly verify their identities to each other through mutual attestation (IETF Datatracker, 2024). It is worth noting that non-technical solutions for verifying chip location, involving the physical inspection of data centers, have also been proposed (Shavit, 2023; Baker, 2023).

Designing hard-to-spoof chip IDs. Secure IDs for chips (either software- or hardware-based) could help with traceability and preventing unauthorized use. For example, unique identifiers and information may be engraved during routine water fabrication flows (MULTIBEAM, 2024), or physical unclonable functions – devices that "[exploit] inherent randomness introduced during manufacturing to give a physical entity a unique 'fingerprint' or trust anchor" (Gao et al., 2020) – may be added. Open questions include understanding the security of these approaches, as well as their feasibility, given potential trade offs between usability, security, and effects on chip performance.

5.2.2 Verification of Compute Workloads

Motivation: It could be valuable for AI developers and deployers to be able to reliably verify their compute usage, for example, which chips were used to train their models and for how long. Likewise, chip owners, such as cloud providers, may want to demonstrate which models were trained on their compute, so as to provide evidence that their compute is not being used for unreported large-scale training runs.[19] Such mechanisms may be of particular utility in international and extra-territorial situations in which levels of trust between verifier and prover may be limited. Any verification scheme would need to uphold privacy of both user data and intellectual property.

We note that this is one area of research that itself is dual-use. While potentially beneficial in the above example cases, some implementations of such mechanisms could enable far-reaching control and monitoring of AI chips. However, careful design of these mechanisms can limit the scope of powers given to a regulator, for example by only requiring data flows for voluntary verification, rather than remote monitoring or control.

Open Problems:

Verifying properties of workloads using TEEs. It may be possible to use TEEs (or other hardware security technologies) to attest to the exact program code and model being run (Chen et al., 2019, see also Section 6.2). For example, TEEs could, given inputs along with a function to evaluate on them, return a signature that accompanies the output from the computation attesting to the computation being run as intended. Alternatively, the chip could also return a hash of the inputs and outputs of the computation. This would allow the prover to keep the inputs and outputs private, instead proving ownership by demonstrating that they can generate the hash returned by the chip during computation. A verifier could also use a TEE to confidentially run arbitrary tests on the model weights or other data. Despite promising theoretical possibilities, further work will be required to be able to implement the above in practice. In particular, firmware to implement the above solution on existing hardware is lacking, and morover, may need to be unusually secure. Advancements will also be needed if solutions such as the above are to be feasible at the largest scales, without introducing prohibitive overhead costs.

Verifying properties of workloads with a trusted neutral cluster. If TEEs are unavailable or impractical, another approach could be to save hashed snapshots of neural network weights during training along with information about the training run being conducted – that is, a *training transcript*. This information may then be used to verify that the training transcript provided would have resulted in the weights, with the use of a trusted neutral cluster (Shavit, 2023). Current challenges to implementing this procedure include

[19] Reporting of training runs above the threshold of 10ˆ26 floating-point operations (FLOP) is required, for example, by the US Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Office of the President, 2023).

difficulties in accounting for randomness in training procedures, building sufficiently trustworthy neutral clusters, and finding efficient methods for proving the authenticity of training transcripts that scale to the largest models (Shavit, 2023).

Verifying compute usage of large, non-AI workloads. Owners or users of large clusters may wish to demonstrate that their clusters were used for a large, non-AI workload (for example, climate simulations), as such use would not fall under the purview of AI regulation. One approach is workload classification, discussed above. There may also be viable computational approaches to verification, potentially including analogues to proof-of-learning methods (Jia et al., 2021), that could be explored in future research.

5.3 Models and Algorithms

Example Research Questions
55. How can model properties be verified with full access to the model? (5.3.1)
56. How can the risk associated with a given context, query, and AI response be assessed in order to obtain assurances about the system's compliance with safety requirements? (5.3.1)
57. What should constitute the lower bar for tracking updates to models, for example in a model registry? (5.3.2)
58. Could proof-of-learning be used to demonstrate and verify model ownership? (5.3.3)
59. How can proof-of-learning mechanisms be made robust to adversarial spoofing? (5.3.3)

5.3.1 Verification of Model Properties

Motivation: In order for system developers or deployers to demonstrate compliance with regulatory requirements, it may be necessary to prove claims regarding model properties and information. Verifiable properties could include model architecture, training procedures, or performance metrics, enabling developers to formally demonstrate compliance with any mandated technical specifications.

Open Problems:

Verifying claimed capabilities and performance characteristics with full model access. Model properties could be verified through formal verification methods if the verifier has full access to the model, as in the case of the model developer. Such methods aim to mathematically prove that a given system can(not) respond in particular ways to particular inputs (Katz et al., 2017b;a; Kuper et al., 2018; Katz et al., 2019). For instance, formal verification was used to study the safety of neural networks used for unmanned aircraft collision avoidance (Irfan et al., 2020). However, such methods remain largely untested for advanced AI models (Dalrymple et al., 2024). In particular, many methods quickly become prohibitively complex when scaled up to contemporary state-of-the-art models. While there exist additional methods for verifying properties such as performance metrics without full access (see Section 5.4.1), research could focus on more efficient methods given full access to the model. Furthermore, verifying properties such as a system's architecture or training procedure remain open questions.

5.3.2 Verification of Dynamic Systems

Motivation: Modern AI systems, such as ChatGPT, are not based on static models. Rather, they consist of multiple models and components, for example, mixture-of-experts, input filters, and output filters, that undergo change throughout their life cycle. This poses an oversight challenge due the ever-changing nature of many systems throughout their deployment life cycle. Having a reliable, accessible process for versioning could help to monitor system updates and their impacts.

Open Problems:

Tracking versioning and updates. Key open questions in this context relate to how model versioning and post-deployment modifications should be kept track of, especially for models that undergo frequent updates. One approach could be to have registries that track models over time, however, it's not clear what information should be stored in such a registry, nor how the information could be verified. Other approaches that can be useful as a starting point to verify dynamic models include reward reports for reinforcement learning (Gilbert et al., 2023), ecosystem graphs (Bommasani et al., 2023c), or instructional fingerprinting of foundation models (Xu et al., 2024).

5.3.3 Proof-of-Learning

Motivation: In the current landscape, there is no mechanism for a model developer to prove that they have invested the computational resources required to train a given model. Such a proof could be used for resolving ownership disputes when models are released or stolen by allowing the developer to attest to their having trained the model (Tramèr et al., 2016; Orekondy et al., 2018; Jia et al., 2021). Additionally, proof-of-learning could aid in defending against accidental or malicious corruption of the training process when performing distributed training across multiple workers (Li et al., 2014; Jia et al., 2021).

Open Problems:

Scalable proof-of-learning. (Jia et al., 2021) were the first to formalize the notion of proof-of-learning for AI models. The authors demonstrated that stochastic gradient descent accumulates "secret information due to its stochasticity," which they show can be used to construct "a proof-of-learning which demonstrates that a party has expended the compute required to obtain a set of model parameters correctly" (Jia et al., 2021). Alternatively, (Goldwasser et al., 2021) develop *Probably Approximately Correct* verification, in which a weak verifier interacts with a strong prover to test whether the model trained by the prover has a low loss relative to the best possible model, with respect to a given loss function. Scaling these techniques such that they remain practical given the growing training compute budgets of foundation models is an open challenge.

Designing adversarially robust proof-of-learning. Since the introduction of proof-of-learning in (Jia et al., 2021), subsequent work has demonstrated its vulnerability to adversarial attacks – that is, false proofs that are cheap for an adversarial prover to generate (Zhang et al., 2022; Fang et al., 2023). In particular, Fang et al. (2023) demonstrate "systemic vulnerabilities of proof-of-learning" and which depend on advances in understanding optimization to be sufficiently addressed. While Choi et al. (2023) suggest a protocol to counter these vulnerabilities through memorization-based tests, and fixing the initialization and data order, they only test their protocol for single attacks and not for composite attacks. Their protocol further only covers language models, and has not been tested for other modalities. Future work could aim to assess these claims, and aim to increase the robustness of proof-of-learning to adversaries.

5.4 Deployment

Example Research Questions
60. How can audit registries be used to provide end-to-end verification along the AI value chain? (5.4.1)
61. How should verification information from model registries be presented to users? (5.4.1)
62. Can zero-knowledge proofs be applied to demonstrate a model's compliance with hypothetical mandated criteria, without directly disclosing architectural details? (5.4.1)
63. How can it be verified that the model version on which an evaluation or audit was performed is the same as is deployed? (5.4.1)
64. How can the implementation of safety measures be verified at deployment? (5.4.1)
65. How can output watermarking schemes be made robust to adversarial attempts at removal? (5.4.2)
66. How can metadata watermarking be applied to AI-generated content? (5.4.2)

67. How robust can AI content detectors be expected to be in light of continuing advances in generative AI? (5.4.2)

68. How should AI-generated content detectors handle cases of genuine images that have been modified or edited with AI tools? (5.4.2)

5.4.1 Verifiable Audits

Motivation: As discussed in Section 3, external audits and assessment have been proposed as crucial components of governance regimes (Raji et al., 2022b; Mökander et al., 2023). Being able to attest to an audit's process and outcome could establish greater trust between model developers, third-party auditors, and governments by proving compliance with regulatory requirements. Trust could also be established with end-users by enabling them to verify that the model with which they are interacting has been shown to have the properties claimed by developers (Godinot et al., 2024), for example in model cards (Mitchell et al., 2019), official communications (for example, Anthropic, 2024a), or technical papers (Gemini Team et al., 2023; OpenAI et al., 2024). Verifying audit results is often made more challenging as a result of access to models often being restricted due to IP and security concerns (South et al., 2024).

Open Problems:

Verifying claimed capabilities and performance characteristics without full model access. Preliminary work has explored how the application of zero-knowledge proofs to AI systems can enable privacy-preserving verifications of claimed system properties, as well as confirmation that model weights used for inference match those on which an audit was run (South et al., 2024; Waiwitlikhit et al., 2024; Sun et al., 2024; Yadav et al., 2024). However, due to the high computational overhead associated with these methods, addressing speed constraints will be necessary if such methods are to be applied to larger models. Current approaches that future work could build on include GPU acceleration (Sun et al., 2024) or proof splitting (South et al., 2024).

Verifying audit results at inference time. In theory, verified computing — such as through TEEs (Sabt et al., 2015), zero-knowledge proofs (Fiege et al., 1987), or secure multi-party computation (Goldreich, 1998) with active security — could facilitate verifiable audits in a two-stage process. In the first step, a model developer could load an inference pipeline[20] into a cluster of enclave computers. Upon an auditor concluding their study of the system (for example, based on the approaches outlined in the previous paragraph and in 5.3.1), they could ask the cluster of enclaves to produce a certificate of the pipeline that was evaluated, which is then stored in a public audit registry. In the second stage, a user, when interacting with the secured pipeline, could request a corresponding certificate with each received generation. Using such a method, this consumer could know that the AI pipeline they're receiving generations from is the same pipeline that was previously evaluated to be safe. If any change was made to the pipeline, the certificates would not match, and the user would know that they're receiving generations from a pipeline which has not been evaluated. However, given the dynamic nature of current models in use (see Section 5.3.2), changes may occur more frequently than audits of such models, which poses an open challenge to this proposal. Another open problem is that this pipeline requires that all evaluation and inference is done in enclaves and with significant computational overhead, effectively limiting verifiable audits for a few critical systems, and necessitating more scalable structures for verifying audits. It will also be necessary to find a way for consumers to be informed of the outputs of this verification in a low-friction way, as in the case of browsers that provide warnings for websites without HTTPS certificates. Finally, it is unclear how secure this method is against attempts to exfiltrate model weights by auditors.

Verifying use of safety measures post-deployment. In safety-critical settings, regulators may want to ensure that safety measures, for example, output filters, are applied to AI models or their outputs (see, for example, Dong et al., 2024a; Leslie et al., 2024; Welbl et al., 2021). Enforcing this may require methods for auditing systems deployed in such domains to check that they do in fact have safeguards that meet these specifications. An open question is how to enforce that additional filters, classifiers, modifications are attached to models deployed in safety-critical domains.

[20] Typically consisting of input pre-processing, model prediction, and post-processing of the model's output.

5.4.2 Verification of AI-generated Content

Motivation: The ability to distinguish between AI-generated and authentic content may be instrumental in verifying the authenticity of information and maintaining public trust in information ecosystems. Stipulations for being able to detect AI-generated media are made in several regulatory efforts, for example in Article 50 of the AI Act (Council of the European Union, 2024) which are, given the state of the art of detection and verification tools, currently unrealizable (Zhang et al., 2023). Methods for verifying AI-generated content can roughly be divided into *ex ante* approaches that mark AI-generated content as such by embedding machine-readable watermarks, and *ex post* methods that aim to classify content as either AI-generated or not, in the absence of a watermark (Ghosal et al., 2023). In addition, watermarks could potentially be used to verify that AI-generated content was created by a particular model, improving accountability by facilitating the identification of responsible parties in case of unintended consequences or misuse.

Open Problems:

Developing robust watermarking schemes. Watermarks – signals placed in output content that are imperceptible to humans, but easily detectable through application of a specific *detector* algorithm – have been proposed as one method for verifying that a particular model generated a given output (Kirchenbauer et al., 2023; Christ et al., 2023; Saberi et al., 2023). However, the level of robustness of watermarks varies between modalities. In particular, the continuous output space of images and audio enables hidden watermarking that is more effective than for text (Ghosal et al., 2023). As such, future work could aim to address the relative lack of robustness of watermarks in the case of AI-generated text (Zhang et al., 2023; Liu et al., 2023a). Additionally, research could aim to address the possibility that watermarks are easy to fake – for example, having two similar models that produce watermarks that cannot be distinguished (Srinivasan, 2024).

Designing robust AI content detectors. While efforts to develop methods for the detection of AI-generated content have seen increased attention in the last two years (Sadasivan et al., 2023; Corvi et al., 2023; Berber Sardinha, 2024), these methods have not always held up to independent evaluation (Weber-Wulff et al., 2023). As generative systems improve, it will be increasingly difficult to develop methods (machine learning-based or otherwise) to distinguish their output from genuine media. Continued work will need to be done in order to improve and maintain the efficacy of AI content detectors in light of this continued advancement.

Utilizing verifiable meta-data to identify authentic content. An alternative to identifying AI-generated content could be to develop ways for a content creator to verify their content as AI-generated or authentic by adding verifiable meta-data to it (Jain et al., 2023c; Knott et al., 2023). For example, the Coalition for Content Provenance and Authenticity (C2PA) is tackling this issue by developing standards for the certification of the provenance of digital content (C2PA, 2022). Similar work is being done by the (Content Authenticity Initiative, 2024). This could be useful for AI labs to label content, but also for creators of non-AI-generated media to label their authentic content as such using the same standard. However, a significant drawback of this approach is that it is not robust to adversaries, as meta-data can easily be stripped from the content – a limitation which future research could aim to address.

Verifying authentic content modified using AI. Complications arise when going beyond the binary distinction of AI-generated content on the one hand, and human-generated on the other. For example, it is currently unclear how AI content detectors should respond to authentic images that have been modified using generative AI tools. Future work could aim to assess the suitability of AI-content generation tools for detecting such cases, or design detectors that are able to distinguish between AI-generated, AI-modified, and authentic content.

6 Security

In this section, we consider security in the context of AI governance, which aims to ensure that unauthorized actors are not able to access systems and infrastructure not intended for their use, nor use systems for malicious purposes. Being able to give security guarantees across system components could be helpful for a number of reasons. Increased security can strengthen a wide array of governance actions through reducing the risk of regulatory requirements being subverted. For example, comprehensive security measures can protect the confidentiality of training data, ensuring that AI systems developed using sensitive personal information remain in compliance with data protection laws.

It should be noted that security is one of the areas of this report that comes closest to topics within AI safety. As such, many of the topics discussed below under the umbrella of TAIG could also be viewed through the framing of improving AI safety, or otherwise be closely related to topics that can. However, due to the reasons above, we decided to include security within this report nonetheless.

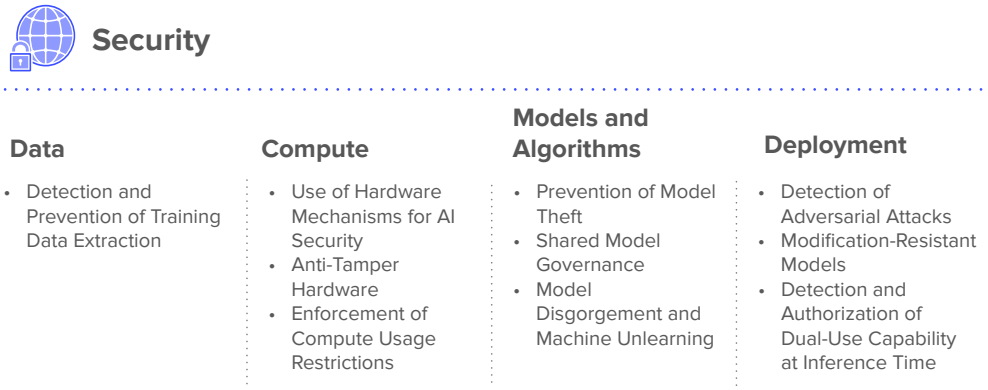


Figure 5: Open problem areas in the *Security* capacity, organized by target

6.1 Data

Example Research Questions
69. How can attempted data extraction attacks be reliably identified? (6.1.1)
70. How can AI systems be made robust to data extraction attacks? (6.1.1)
71. How can methods for restricting *verbatim* reproduction of training data be generalized to protect the same information being extracted in a slightly different form? (6.1.1)

6.1.1 Detection and Prevention of Training Data Extraction

Motivation: Prior research has demonstrated how large amounts of models' training data can be extracted *verbatim*, with a variety of methods applicable in both black- and white-box settings (Carlini et al., 2023c; Nasr et al., 2023; Shi et al., 2023; Balle et al., 2022; Carlini et al., 2019; 2021; Duan et al., 2024b; Prashanth et al., 2024). Short of building models that are robust to extraction attacks, having the ability to detect them could enable API-level defenses that can block model outputs upon detection of a potential attack, or the introduction and enforcement of litigation against perpetrators of extraction attacks.

Open Problems:

Improving system robustness to extraction attacks. De-duplication of training data has been shown to assist in reducing memorization, and hence extraction, of specific data-points (Kandpal et al., 2022), though

Nasr et al. (2023) suggest that this provides only marginal improvement. Alternatively, there may be *post-hoc* interventions that do not alter a model's memorization of its training data, but nonetheless decrease its propensity to reproduce it, with one such example being machine unlearning (see Section 6.3.3). However, acute challenges still remain. For example, restricting the *verbatim* reproduction of training samples does not prevent the same information being generated by models with slight rewording or reformatting (Ippolito et al., 2023). Furthermore, guarding against the reproduction of specific samples has been found to expose previously-safe samples to the same attacks – a phenomenon dubbed the "Privacy Onion Effect" (Carlini et al., 2022). Finally, a related yet under-explored area of concern regards the potential for large-scale data extraction from retrieval datasets (Qi et al., 2024c).

Detecting attempted data extraction attacks. Proposed methods for detecting potential data extraction attacks are noticeably absent from the literature, with most publications on this topic aiming to identify model vulnerabilities to such attacks. Potential methods for detecting extraction attacks could focus on either the model inputs or outputs, allowing model providers to filter out suspicious prompts, or outputs that bear a close resemblance to training samples, respectively.

6.2 Compute

Example Research Questions

72. How can hardware-enabled governance methods be implemented at scale to ensure the security of a compute cluster? (6.2.1)
73. How can it be ensured that given code, along with model weights, can only be executed with a license that is verified on-chip, so that distributed AI executables can only be run on approved chips? (6.2.1)
74. How can on-chip governance firmware be modified or updated while the chip is in operation while remaining resistant to potential attacks? (6.2.1)
75. How secure are existing implementations of TEEs on AI accelerators? (6.2.1)
76. How can methods for tamper-evidence or responsiveness be reconciled with the performance demands of high-end AI accelerators? (6.2.2)
77. How secure are existing approaches to tamper-evidence and responsiveness? (6.2.2)
78. How can tamper-proofing methods incorporate self-destruct mechanisms in case of attempted tampering? (6.2.2)
79. How could the use of high-end chips in training foundation models be prevented? (6.2.3)
80. Can enforceable mechanisms be developed to allow for the export of chips under predefined conditions? (6.2.3)

6.2.1 Use of Hardware Mechanisms for AI Security

Motivation: The integration of hardware mechanisms such as TEEs into AI computing clusters could ensure the confidentiality and integrity of workloads (Li et al., 2023b; Geppert et al., 2022; Mo et al., 2024) while also greatly aiding with AI security and attestation (Nevo et al., 2024; Kulp et al., 2024; Aarne et al., 2024). This in turn would assist in implementing many of the aforementioned problem areas relating to verification and access.

Open problems:

Ensuring utility of TEEs for hardware-enabled governance and security. While TEEs have seen broad adoption on CPUs,[21] application to AI accelerators (such as GPUs and TPUs) has so far been limited. The most notable example is Nvidia's incorporation of a TEE in its H100 GPU, referring to it as "NVIDIA Confidential Computing" (Dhanuskodi et al., 2023; Hande, 2023; Apsey et al., 2023). However, the H100 implementation "still may not support all of the mechanisms required for an ideal implementation of [hardware governance measures]" (Aarne et al., 2024). In particular, questions remain regarding whether TEEs can robustly attest to the identity of a specific chip or the data that it is processing. Furthermore, current implementations mostly do not support confidential computing across multiple individual accelerators in a cluster. Further work could investigate the extent to which such functions are supported by current and next-generation chips, or aim to specify hardware, firmware, or software requirements necessary for robustly implementing such functions at the scale of compute clusters, or even entire data centers.

Ensuring security of TEEs on AI accelerators. Given the novelty of the application of TEEs to high-end, AI-specific hardware, it is as-yet unknown how secure such systems are in practice due to a lack of independent testing. Previous independent testing of CPU TEEs has uncovered numerous potential vulnerabilities (Muñoz et al., 2023; van Schaik et al., 2022). Additional security research into GPU TEEs and other security features that they rely on, such as Nvidia's GPU system processor, would be valuable for identifying areas of improvement and assessing whether these features can be relied on for different governance applications.

6.2.2 Anti-Tamper Hardware

Motivation: Some of the aforementioned hardware-enabled governance mechanisms, such as verifying compute workloads (Section 5.2), rely on the assumption that the hardware in question has not been compromised or tampered with, for example, if dependent on a TEE for implementation. However, well-resourced adversaries may aim to physically tamper with chips in order to obviate the defensive protections installed on them. It could therefore be useful to disincentivize or prevent such tampering. This can come in the form of *tamper evidence*, whereby physical manipulation of the chip is detectable after the fact, or through *tamper responsiveness*, whereby tampering with the chip triggers an automatic response ranging from deletion of sensitive information stored on the chip, known as *zeroization*, to permanent self-destruction.

Open problems:

Reconciling tamper evidence and responsiveness with practical requirements of state-of-the-art AI hardware. While approaches to tamper evidence and responsiveness have been proposed (Immler et al., 2018), ensuring that such approaches are compatible with the unique requirements of state-of-the-art AI accelerators, while retaining affordability and scalability, is an outstanding challenge (Aarne et al., 2024). For example, demanding cooling requirements and high-bandwidth interconnect between chips pose a challenge due to the need to bridge the interior and exterior of the tamper-proof enclosure (Obermaier & Immler, 2018).

Ensuring the robustness of anti-tamper approaches. Established approaches to tamper evidence and responsiveness have depended on the use of specialized packaging[22] that encases the chip and is unable to be removed without leaving visible traces of damage. In the case of tamper responsiveness, the packaging may carry an electric current such that damage disturbs the current, acting as a trigger for zeroization or other active responses. More advanced methods use physical unclonable functions – a method that "exploits inherent randomness introduced during manufacturing to give a physical entity a unique 'fingerprint' or trust anchor" (Gao et al., 2020) – to remotely attest that a chip has not been tampered with (Immler et al., 2018; 2019; Obermaier & Immler, 2018). However, evidence pertaining to the practical success of tamper-proofing has been limited (Aarne et al., 2024; Kulp et al., 2024). Further research is needed if we are to increase confidence in anti-tamper measures when applied to AI hardware security.

[21] See, for example, specifications by Intel (Intel, 2022) and AMD (AMD, 2024).

[22] In this context, *packaging* refers specifically to a physical security enclosure that encases a hardware device as opposed to packaging, such as a cardboard box or other container, in which the device may be stored or transported.

6.2.3 Enforcement of Compute Usage Restrictions

Motivation: Recent attention in compute governance has been paid to export controls placed on cutting-edge chips of the type used in large-scale training of AI systems (Bureau of Industry and Security, 2022a; Allen, 2022). However, export controls are a blunt instrument with a high potential for collateral damage by restricting the sale of affected chips for legitimate uses. Indeed, the Bureau of Industry and Security, the US executive agency responsible for such export controls, itself put a call out for "public comments on proposed technical solutions that limit items specified under [the export controls] from being used in conjunction with large numbers of other such items in ways that enable training large dual-use AI foundation models with capabilities of concern" (Bureau of Industry and Security, 2023), presumably acknowledging that technological developments for disentangling legitimate and malicious uses of high-end chips would be desirable. It is worth noting, however, that there is considerable disagreement over both the viability of such solutions, with concerns raised regarding the level of confidentiality of such measures, as well as the possibility of their being circumvented (Ting-Fang, 2023; Patel, 2023; Grunewald & Aird, 2023; Fist & Grunewald, 2023).

Open problems:

Implementing remote attestation for disaggregated machines. It would be useful to verify the particular set of hardware components – in this case, AI chips – that are part of the same cluster. This would assist with hardware-based methods for verifying properties of workloads, which typically rely on knowing which chips are participating in a workload, and for ensuring that end-users are complying with export control obligations. An open question is how remote attestation could work for disaggregated machines (Google Cloud, 2024) or for heterogeneous devices. These mechanisms could allow the nature and acceptability of the configuration to be remotely attested to. Other relevant projects in this context that may serve as starting points are *Caliptra* and *OpenTitan*.

Restricting particular cluster configurations. It may also be possible to restrict possible cluster configurations to assist with export control policies. One proposed approach involves restricting the communication bandwidth between GPUs to prevent "many consumer device-chips from being aggregated into a supercomputer" (Kulp et al., 2024). It may be possible to build such a system on top of existing features such as trusted platform modules (Hosseinzadeh et al., 2019), or it may require new protocols and new hardware-level features – all of which are open problems.

6.3 Models and Algorithms

Example Research Questions

81. What cybersecurity measures can be taken at the infrastructure level to protect model weights from theft by an adversary? (6.3.1)
82. How can models be protected from inference attacks aiming to reproduce or replicate model weights and architecture? (6.3.1)
83. What are the most promising methods for enabling shared model governance? (6.3.2)
84. How should the success of different model unlearning techniques be evaluated? (6.3.3)
85. How can it be ensured that machine unlearning and model editing techniques do not cause unwanted side-effects such as removing concepts that were not explicitly targeted? (6.3.3)
86. How effective are model unlearning and model editing techniques when applied to multi-lingual or multi-modal models? (6.3.3)

6.3.1 Prevention of Model Theft

Motivation: As models become more capable they could become an increasingly valuable target for theft by adversarial parties wanting to put them to their own potential (mis)use. Similarly, as state-of-the-art models become more broadly integrated into the economy and society, the attack surface will increase, potentially leading to a greater threat of exfiltration (Nevo et al., 2024). It follows that securing model weights, and other system components, might become an increasing priority to prevent theft or model access by unauthorized parties that may undermine governance initiatives aimed at ensuring customer safety and national security (Nevo et al., 2024).

Open problems:

Ensuring adequate cybersecurity for model weights. Protecting model weights against exfiltration attempts requires protections against insider and outsider threats (Nevo et al., 2024). This includes standards for physical security of the data center facility itself, as well as of the hardware and software stacks (OpenAI, 2024b).[23] Improved coordination between actors facing similar threats might also assist defenders in understanding the threat landscape and better protecting their assets during training and deployment. Further analysis of potential threat vectors, as well as development of physical and cybersecurity measures including and beyond those in (Nevo et al., 2024), would help to identify and address these risks.

Defending against model inference attacks. Alternatively, adversaries may try to extract or replicate models through attacks to a query API (Orekondy et al., 2018; Tramèr et al., 2016; Jagielski et al., 2020; Carlini et al., 2020; 2024), logit values (Carlini et al., 2024) or side-channel attacks (Wei et al., 2020). Further research could aim to quantify threats and develop methods for defending against these, and other, forms of model extraction attacks.

6.3.2 Shared Model Governance

Motivation: Shared model governance refers to the practice of distributing control over a model's training or inference across multiple parties, such that training or inference can only be carried out with the agreement of all parties (Bluemke et al., 2023). The ability to distribute control of a model in this way could have many potential use cases – for example, if multiple diverse actors want to pool investment for training a shared model where each actor has specific requirements for how the model is trained. This may also be applicable in the case of international collaboration between state-backed institutes wanting to collaborate on AI research (Ho et al., 2023).

Open problems:

Enabling shared model governance through model splitting. One proposed approach for technically-enforced shared model governance is *model splitting* (Martic et al., 2018) – that is, "distributing a deep learning model between multiple parties such that each party holds a disjoint subset of the model's parameters." Martic et al. (2018) also investigate the resulting question of how computationally expensive it would be for a single actor to reconstruct the entire model, starting from their share of the parameters. A similar approach is taken by *SplitNN* (Vepakomma et al., 2018; Ceballos et al., 2020), though the emphasis is placed on model splitting to achieve data privacy, rather than shared model governance. Given the relatively small amount of prior work on model splitting for shared model governance, future work could aim to provide further proofs of concept and evaluate their efficacy.

Enabling shared model governance through secure multi-party computation and homomorphic encryption. Two potential alternative approaches for achieving shared model governance are applying either secure multi-party computation (SMPC) (Yao, 1982; 1986; Evans et al., 2018), or homomorphic encryption (HE) (Gentry, 2009; Acar et al., 2018). Though usually applied to AI for the purposes of data privacy (Knott et al., 2021; Kumar et al., 2019; Tan et al., 2021; Guo et al., 2022; Riazi et al., 2018), or model privacy (Trask, 2017; Dahl, 2017; Ryffel et al., 2018), both SMPC and HE could potentially be leveraged for shared model governance. For example, using SMPC, a model creator could take each parameter within a model and split it into multiple *shares*, distributing such shares across *shareholders*. Alternatively, using HE, a model could be encrypted using one or more private keys such that the ability to decrypt model results relies upon

[23] Data center security standards: (International Organization for Standardization, 2021; Wikipedia contributors, 2023).

the application of the private key of all parties. However, both HE and SMPC have performance concerns, with even state-of-the-art encrypted deep learning approaches yielding 100x performance slowdown (Wagh et al., 2019; 2020; Stoian et al., 2023; Frery et al., 2023). Future work could investigate this potential in more detail, aiming to provide proof-of-concept demonstrations of how shared model governance using either HE or SMPC could be achieved. Alternatively, future research could aim to reduce the high computational overheads of HE and SMPC.

Enabling shared model governance through TEEs. Finally, shared governance could potentially be achieved through the application of TEEs. In this case, multiple parties could upload a program to a TEE, with the chip providing evidence on the software program being run. With this evidence, all parties can know how any information they upload to the enclave will be handled. TEEs may also be able to reinforce the above approaches, for example, SMPC. Future research could aim to provide proof-of-concept demonstrations of shared model governance through the use of TEEs given the current hypothetical nature of this approach.

6.3.3 Model Disgorgement and Machine Unlearning

Motivation: The concepts of *model disgorgement* (Achille et al., 2024) and *machine unlearning* (Bourtoule et al., 2021; Nguyen et al., 2022; Shaik et al., 2023; Si et al., 2023; Eldan & Russinovich, 2023; Yao et al., 2023; Liu et al., 2024a;b; Goel et al., 2024) have been proposed as methods for removing memorized information or otherwise nullifying the impact of a model's having been trained on problematic data. This could potentially introduce a pathway through which harms of reproducing inappropriate or copyright data could be addressed in cases where action was not taken during data curation or model training. Related methods for direct model editing (Mitchell et al., 2021; 2022a; Meng et al., 2022; Hernandez et al., 2023) to remove learned harmful concepts, through editing activations (Zou et al., 2023a; Turner et al., 2023), concept erasure (Ravfogel et al., 2022; Belrose et al., 2023), or targeted lesions (Li et al., 2023a; Wu et al., 2023), could provide alternative approaches to achieving these aims.

Open problems:

Ensuring unlearning methods are robust and well-calibrated. Machine unlearning involves an interplay between specificity of, and generalization from, concepts to be unlearned. In particular, methods that successfully generalize can aid in cases where the unlearning target is hard to precisely specify. However, generalization may open the door to unintended side-effects if it results in the removal of non-target concepts (Cohen et al., 2024b). A challenge to be addressed then is to ensure that methods for machine unlearning and model disgorgement are well-calibrated in that they successfully generalize to comprehensively remove target concepts, while avoiding the removal of benign concepts.

Extending unlearning and model editing to cross-lingual and cross-modal models. As trends towards multilingual (Üstün et al., 2024) and multi-modal (Yin et al., 2023) models continue, there will be a need to extend model unlearning and editing techniques to these models. Questions remain as to the efficacy of such techniques when applied in such cases (Si et al., 2023), for example, regarding whether models retain concepts in other languages, despite that concept having been unlearned in English.

Evaluating the efficacy of unlearning and direct model editing techniques. A further outstanding question is how the efficacy of unlearning attempts can be evaluated (Lynch et al., 2024; Shi et al., 2024). Evaluations should aim to assess not only whether the influence of the specified unlearning targets has indeed been removed and that model performance in other domains has not been adversely affected (Li et al., 2024b), but also identify potential *ripple effects* that may have resulted from an application of unlearning or model editing (Cohen et al., 2024b).

6.4 Deployment

Example Research Questions
87. How can the robustness of methods for detecting adversarial attacks be improved? (6.4.1)
88. What interventions are most effective for handling detected adversarial attacks at inference time? (6.4.1)
89. How can a model be made resistant to being fine-tuned for malicious tasks, while still allowing for benign fine-tuning? (6.4.2)
90. How can the request of dual-use system capabilities be reliably detected? (6.4.3)
91. How could authorization of user identity be used as a gate for dual-use model capabilities? (6.4.3)

6.4.1 Detection of Adversarial Attacks

Motivation: Adversarial attacks refer to deliberate attempts to exploit inherent model vulnerabilities in order to make the model behave incorrectly or harmfully, for example, by outputting offensive or toxic language (Lohn, 2020; Shayegani et al., 2023; Vassilev et al., 2024). In the case of language models, such attacks often take the form of modifying the user input so as to bypass any implemented safety filters. Some attacks are transferable across different models (Zou et al., 2023b) and defenses against adversarial attacks are typically narrow and brittle (Narayanan & Kapoor, 2024). The ability to detect such attacks could enable the application of targeted system-level defenses, such as halting or filtering system output, separate from relying solely on the underlying model's robustness to attacks. Furthermore, having empirical evidence on the frequency of attacks can help inform deployment corrections (Section 7.2) and threat models (Section 8.1).

While some system-level defenses against adversarial attacks exist, it is important to note that many such protective measures can only be implemented effectively in an application or deployment context (Narayanan & Kapoor, 2024). Though directly improving the adversarial robustness at the model level is a related active research area (Vassilev et al., 2024; Hu et al., 2024; Su et al., 2024) here we emphasize the closely related issue of being able to detect and handle potential adversarial attacks at inference time due to its relevance for governance interventions as mentioned above.

Open problems:

Detecting adversarial inputs and outputs. Being able to detect and classify user inputs to a model as potential adversarial attacks allows for filtering (Jain et al., 2023a; Aldahdooh et al., 2022) or preprocessing (Cohen et al., 2019; Nie et al., 2022; Kumar et al., 2023; Jain et al., 2023a; Zhou et al., 2024) of concerning inputs before being given to the model. Alternatively, model outputs could be filtered with the aim of detecting a model's response to adversarial attacks in order to remove them before reaching the user (Phute et al., 2024; Greenblatt et al., 2023). Current techniques for detection, however, can suffer from a lack of robustness themselves or may introduce significant latency for the user (Glukhov et al., 2024).

6.4.2 Modification-Resistant Models

Motivation: Post-deployment fine-tuning is a common method for user customization of language models, either through an API or locally in the case of downloadable models. Fine-tuning, and other post-training enhancements have been theorized to have an outsized impact on downstream performance (Davidson et al., 2023). However, just as fine-tuning can be used to customize a model for legitimate and beneficial use cases, it can just as easily be used to adapt a model for malicious purposes, often with small amounts of data (Jain et al., 2023b; Yang et al., 2023; Qi et al., 2023; Lermen et al., 2023; Zhan et al., 2024; Qi et al., 2024a;b). Fine-tuning attacks often aim to achieve similar aims to those of adversarial attacks above, though through customizing models directly, rather than by modifying input prompts. Having methods for preventing the customization of models for malicious use could reduce misuse risks associated with open-weight release,

thus expanding the range of potential deployment options and promoting the numerous benefits of more open release strategies.

Open problems:

Preventing the modification of models for malicious tasks. An open question is whether there exist technical methods that restrict a model's amenability to being fine-tuned (or modified through other methods) for harmful uses, while retaining the ability to be modified for benign uses (Rosati et al., 2024b; Peng et al., 2024). Potential methods may aim to raise the computational cost of fine-tuning on harmful data to prohibitive levels (Henderson et al., 2023b; Deng et al., 2024; Rosati et al., 2024a; Tamirisa et al., 2024) or make models resistant to learning from harmful data (Zhou et al., 2023b; Huang et al., 2024c; Hsu et al., 2024; Huang et al., 2024b). However, given the nascency of these techniques, future research could aim to establish their robustness in practice.

6.4.3 Detection and Authorization of Dual-Use Capability at Inference Time

Motivation: In the event that model assessments have flagged a system's competence in dual-use domains – that is, domains which can have both beneficial and harmful applications – model providers might need to avoid exposing these capabilities publicly by default in order to avoid misuse. However, completely removing these capabilities may not be feasible, or economically favorable due to the legitimate and beneficial use-cases, such as a cybersecurity professional using a system to aid in the identification and patching of software vulnerabilities. This differs from the discussion of adversarial attacks above, for which the aim of an attack is taken to be unambiguously harmful.

Open Problems:

Detecting requests of dual-use capabilities. Guarding against malicious uses of dual-use capabilities is currently imperfectly achieved by conducting safety fine-tuning so that the model refuses to respond to malicious requests. However, this approach is not robust to jailbreaks that pose as legitimate requests for such capabilities (Wei et al., 2023; Fang et al., 2024). An alternative approach could be to detect all requests for dual-use capabilities. This would allow for the application of separate methods for distinguishing between legitimate and malicious requests, such as independent classifiers trained to separate between legitimate and malicious intent, which may perform better than broad safety fine-tuning.

Requiring authorization for dual-use capabilities. Alternatively, authentication, for example as a certified cybersecurity expert, before accessing certain capabilities may be one way of managing the dual-use nature of general models. This may also have uses for allowing red-teamers or researchers to access such capabilities for research purposes (Longpre et al., 2024a). Such proposals are currently hypothetical, and so future work could aim to propose proof-of-concept demonstrations of how such an authorization scheme could be implemented in practice.

7 Operationalization

Previous sections have discussed concrete technical problems in relation to specific targets in the taxonomy. In contrast, the following two sections will discuss capacities that span across these targets, namely, *operationalization*, and *ecosystem monitoring*.

Operationalization entails the translation of ethical principles, legal requirements, and governance objectives into concrete strategies, procedures, and technical standards. It can also involve the harmonization of terminology and concepts across governance frameworks, for example, NIST's *Risk Management Framework Crosswalks* (NIST, 2023a). Without technical expertise, operationalization efforts may fail to capture the nuances and realities of AI systems, leading to ineffective or even counterproductive governance initiatives.

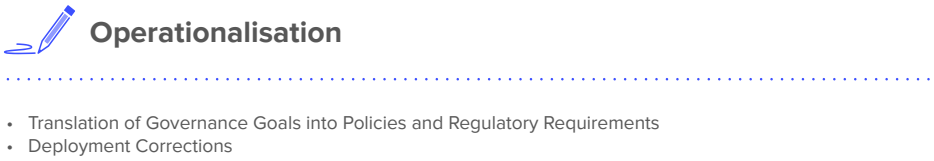


Figure 6: Open problem areas in the *Operationalization* capacity

Example Research Questions
92. What system properties (if any) are the most reliable indicators of risk, and thus candidates for serving as regulatory targets? (7.1)
93. How can AI safety, reliability, and other technical requirements be standardized given an insufficient explanatory understanding of model behavior? (7.1)
94. What are general intervention and correction options if flaws with a model are identified post deployment? (7.2)

7.1 Translation of Governance Goals into Policies and Requirements

Motivation: Policies are often formulated with specific aims in mind, for example, to protect consumer safety, promote fairness, or ensure accountability. In cases of *rule-based regulation*, these aims must be translated into rules that "typically prescribe or prohibit a specific behavior" (Schuett et al., 2024). In many cases this act of translation will demand involvement of technical expertise to provide guidance on the feasibility of proposed rules, as well as the extent to which they achieve a policy's stated aims. For example, the goal of ensuring consumer safety may motivate the introduction of a licensing regime that mandates pre-market safety evaluations. Given the current robustness of such evaluations (see Section 3), this may not only fail to ensure safe and reliable products, but also create a false sense of security (Reuel et al., 2024b; Wu, 2023). For many governance efforts, concrete translations of goals into effective requirements and standards are still lacking (Guha et al., 2024; Pouget, 2023).

Open Problems:

Identifying target dimensions for regulation. Identifying technical dimensions that best align with governance priorities is an open challenge. For example, the current practice of using training compute[24] as a measure of risk may not always be suitable, given that smaller models can outperform larger ones with targeted training. Furthermore, modern AI systems are often the result of multiple data curation and training processes, and it is unclear whether compute expenditure from auxiliary processes should contribute towards

[24]Measured in *floating-point operations* (FLOP).

the final FLOP count. It's also unclear how measures of training compute should take into account techniques such as quantization and drop-out (Hooker, 2024).[25] Aside from training compute, are there other, more precise ways of defining which systems should be subject to regulation? How could such measures account for improved algorithms, and ways in which relevant capabilities can be increased after training (Davidson et al., 2023; Scharre, 2024)?

Detailing and creating standards across the AI system life cycle. While recent AI standard-setting efforts, such as the NIST Risk Management Framework (NIST, 2023b; 2024) and ISO guidance on AI risk management (International Organization for Standardization, 2023), provide valuable general principles, they often lack the technical specificity required for objective assessment of AI systems' compliance with safety and ethical requirements (Pouget, 2023). Additionally, though standard-setting bodies such as IETF, IEEE, ISO, and CEN-CENELEC, along with initiatives such as the Partnership on AI (Partnership on AI, 2023), are working to develop more detailed and verifiable guidance, many areas still require further technical expertise and research (Barrett et al., 2023). One such area is that of security-by-design for hardware, where further work is needed to implement standards across firms, including standards for multi-device attestation in a cluster and TEEs (Cybersecurity and Infrastructure Security Agency, 2023; Kelly et al., 2022, see also Section 6.2). Other goals, such as fairness, can be measured in various ways, and it remains unclear which metrics are most appropriate and effective in which contexts (Parraga et al., 2023; Caton & Haas, 2024; Chouldechova, 2017; Kleinberg, 2018). Finally, standardizing reporting for AI systems, such as the information included in model cards (Mitchell et al., 2019) or data sheets (Gebru et al., 2021), could increase the utility of such practices in governance contexts, motivating the question of what specific information should be included in standardized reports (Kolt et al., 2024; Bommasani et al., 2024).

7.2 Deployment Corrections

Motivation: In the event that flaws are identified in a deployed model, it would be beneficial to adequately respond to the identified risk. Such a scenario could occur either through the identification of previously unobserved capabilities in a deployed model, or through post-training enhancements such as fine-tuning (Davidson et al., 2023). O'Brien et al. (2023) refer to post-deployment responsive actions as "deployment corrections." While they explore this issue from an institutional perspective, for example providing recommendations on how this could be addressed through corporate governance structures and procedures, we see scope for much greater exploration from a technical perspective.

Open problems:

Navigating the continuum of model corrections and interventions. O'Brien et al. (2023) define five categories for deployment corrections: *user-based restrictions*, *access frequency limits*, *capability or feature restrictions*, *use case restrictions*, and *model shutdown*. Within each of these categories there are open questions regarding the feasibility of implementation. For example, *model shutdown* is a relatively extreme action to take upon discovery of a system flaw, and if carried out naïvely could risk major disruption to users, clients, and services that depend on the system in question. Thus, it would be beneficial to have methods in place for minimizing disruption to downstream services in the event that model shutdown is deemed necessary. Furthermore, model shutdown, along with deployment corrections that modify the underlying model, are at odds with the issue of providing model stability and backward-compatibility – features that are particularly relevant for ensuring the reproducibility and replicability of AI research (see Sections 4.3.1 and 5.3.2).

[25] Some recommendations are put forth in (Frontier Model Forum, 2024).

8 Ecosystem Monitoring

Due to the rapid pace of advancements in AI, coupled with uncertainty about future developments, AI governance needs to be forward-looking, future-proof, and adaptive (Guihot et al., 2017; Kolt, 2023; Reuel & Undheim, 2024). To fulfill this goal, decision-makers need to be aware of the multiplicity of stakeholders in the AI ecosystem and how they relate to each other, as well as general trends and potential impacts of current and future AI systems (Wansley, 2016; Ada Lovelace Institute, 2023; Whittlestone & Clark, 2021; Epoch, 2023). Collating and providing such information, which we refer to as *ecosystem monitoring*, can enable AI governance actors to make more informed decisions, better anticipate future challenges, and identify key leverage points for effective governance interventions.

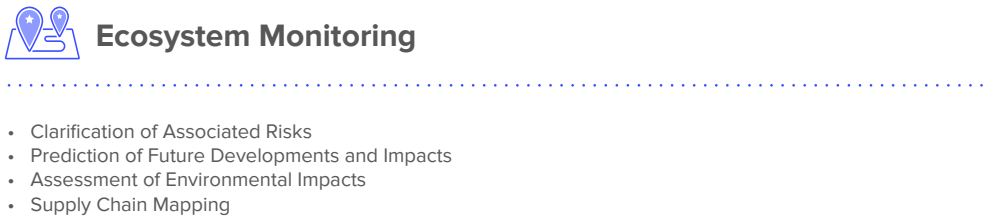


Figure 7: Open problem areas in the *Ecosystem Monitoring* capacity

Example Research Questions
95. What risks, whether from intended or unintended harm, are associated with different (types of) systems? (8.1)
96. How do potential risks differ across domains? (8.1)
97. How can trends and/or properties observed in current systems be extrapolated to make predictions about future systems? (8.2)
98. How could developments in AI-specific hardware impact the governability of compute? (8.2)
99. What information about systems is needed to accurately assess the environmental impact of its development and deployment? (8.3)
100. Given the required information, how can the environmental impact of an AI system be accurately assessed? (8.3)
101. What technical methods can be implemented to create an auditable log of all actors and their contributions throughout the AI development process, from data collection to model deployment? (8.4)

8.1 Clarification of Associated Risks

Motivation: Understanding risks associated with the development and deployment of AI systems enables policymakers to prioritize governance efforts, allocate resources effectively, and determine the urgency of addressing specific risks (Whittlestone & Clark, 2021; Clark, 2023).

Open Problems:

Developing better threat models for risks of AI. While much prior work has intended to lay out taxonomies of risks and harms posed or exacerbated by AI systems (Critch & Russell, 2023; Hendrycks et al., 2023; Weidinger et al., 2022; Abercrombie et al., 2024; Hammond et al., forthcoming; Zeng et al., 2024;

OECD, 2023; Hoffmann & Frase, 2023; Turchin & Denkenberger, 2018; Grabb et al., 2024), detailed threat models have been relatively underexplored. One option for future research could aim to apply standardized risk management approaches, such as causal mapping (Eden et al., 2004; Ackermann et al., 2014), to gain greater clarity into if and how harms from AI may materialize, and where policy could intervene.

Improving incident reporting and monitoring. Additionally, developing improved systems for monitoring and reporting previous or ongoing incidents could not only allow for a more targeted response to ongoing harms, but also facilitate the identification of early warning signals for potential harms (Shane, 2024). AI incident databases have been developed by both the OECD and Partnership on AI, both of which log news articles detailing AI-related incidents (OECD.AI Policy Observatory, 2024; McGregor, 2020). Given that these databases rely solely on public sources, it is likely that only a subset of all incidents are included. In addition, they do not record all details about an incident such as model specifics or deployed guardrails, limiting the utility for analysis of what may have caused an incident. Open questions thus concern how non-public incidents can be reliably reported, as well as what technical information should be reported in order to facilitate meaningful analysis of incidents.

8.2 Prediction of Future Developments and Impacts

Motivation: Anticipating the trajectory and potential impact of AI systems may allow policymakers to proactively set governance priorities, determine the urgency of addressing specific issues, and allocate resources accordingly (Toner et al., 2023). Greater foresight would enable more adaptive and anticipatory approaches to AI governance, which are essential given the rapid pace of AI development (Reuel & Undheim, 2024).

Open Problems:

Measuring and extrapolating from empirical trends. Existing work has aimed to empirically measure trends in training compute (Sevilla et al., 2022) and algorithmic progress (Ho et al., 2024), among others (Epoch, 2023). Future work could aim to extend this effort by quantifying other trends that have not yet been addressed, such as usage patterns of AI in different industries, or assessing the accuracy of predictions based on the extrapolation of observed trends.

Estimating a system's impact before deployment. Estimating the impact of an AI system before deployment, including economic impacts (Eloundou et al., 2024), could help prioritize governance efforts. While such predictions could be aided by more developed threat models (see Section 8.1 above), research may also benefit from technical tools to safely and ethically experiment and simulate potential outcomes without causing harm.

8.3 Assessment of Environmental Impacts

Motivation: The environmental impact of AI systems extends across the entire AI life cycle (Metcalf et al., 2021; Luccioni et al., 2023b; Rakova & Dobbe, 2023), including both during training (Strubell et al., 2019; Patterson et al., 2022) and inference (Luccioni et al., 2023a). Having an accurate understanding of the end-to-end environmental impacts is crucial for policy initiatives, for example, to determine suitable incentives and penalties for encouraging AI developers to reduce the environmental costs associated with their systems.

Open Problems:

Assessing the energy usage of training and hosting systems. Open problems remain due to the logistical challenges of tracking energy consumption and carbon emissions across numerous dynamic system instances. Furthermore, current efforts struggle to take into account energy sources – a factor which can massively affect the overall impact assessment. Ongoing work aims to develop energy ratings for combinations of models and tasks, allowing users to make informed decisions about their system usage, taking into account the environmental impacts of their choice (Luccioni, 2024). Alternatively, tools such as *CodeCarbon*[26] provide developers with real-time estimates for the carbon emissions from running their code. Other work has focused on comparing compute cost on smartphones vs. the cloud (Patterson et al., 2024), best practices for training

[26] https://codecarbon.io/

models (Patterson et al., 2022) and comparing cost for different models (Luccioni et al., 2023a; Luccioni & Hernandez-Garcia, 2023).

Assessing environmental costs of raw resources for building and running data centers. Along with energy, environmental costs may come from other sources along the semiconductor and AI supply chains, for example, from mining and refining the rare earth minerals required for the manufacturing of semiconductors (Kuo et al., 2022; Ruberti, 2023). Additionally, large data centers used for training and hosting AI systems require great quantities of water as part of their cooling systems (Mytton, 2021). Future research could aim to provide in-depth end-to-end predictions of the environmental costs of constructing and maintaining data centers to inform policies aimed at reducing associated environmental impacts.

8.4 Supply Chain Mapping

Motivation: Mapping the AI supply chains can allow policymakers to better understand the complex ecosystem involved in the development and deployment of AI systems. By identifying key actors and processes at each stage of the supply chain, policymakers can target interventions at the most suitable point in the supply chain. Furthermore, existing export controls limiting chip exports to Russia and China have been marked by substantial enforcement difficulties (Allen et al., 2022), and analyses have suggested that AI chips are also likely to become targets for substantial smuggling operations (Grunewald & Aird, 2023; Fist & Grunewald, 2023). By understanding the flow of these resources, authorities can better combat the smuggling of chips and other hardware components.

Open Problems:

Identifying supply chain components and actors. Another area requiring further technical expertise is the identification and assessment of supply chain components. For example, in the context of liability and copyright law, tracking components and design choices made by different actors along the AI supply chain could enable courts to make more precise assessments of potential infringement responsibility (Lee et al., 2024; Longpre et al., 2024b). This granular understanding might be necessary as infringement can occur at multiple points: during data collection, model training, or output generation. If a model produces content resembling copyrighted material, determining liability may require tracing back through the supply chain to identify the source of infringement, whether in training data, model architecture, or generation prompt.

9 Conclusion

In this paper, we presented a broad overview of open problems in technical AI governance (TAIG) across six capacities and four governance targets. We provided a definition of TAIG, a corresponding taxonomy of the work that it entails, and an overview of sub-problems for each sub-area defined in our taxonomy. We also provided links to relevant literature and example research questions that technical researchers could tackle to help advance AI governance efforts.

We also showed that TAIG is not a monolithic problem – instead, it comprises interrelated sub-problems requiring distinct yet complementary approaches. Our taxonomy provides a structured framework for addressing such issues: Each open problem area presents unique technical bottlenecks, such as challenges in attributing model behaviors to training data, ensuring privacy-preserving access, and verifying system properties. By outlining these constraints, we highlight where governance aspirations are hindered by technical limitations and where technical research can drive progress in AI governance.

Our work further highlights that technical work does not exist in isolation. Instead, addressing many of these problems will require cross-disciplinary collaboration among technical researchers, policymakers, industry stakeholders, and civil society. No single discipline can fully resolve AI governance challenges, necessitating integrated expertise across computer science, law, political science, ethics, and economics. As AI systems become more capable and widely deployed, governance strategies must develop alongside technical advances, making TAIG an evolving field and ongoing priority.

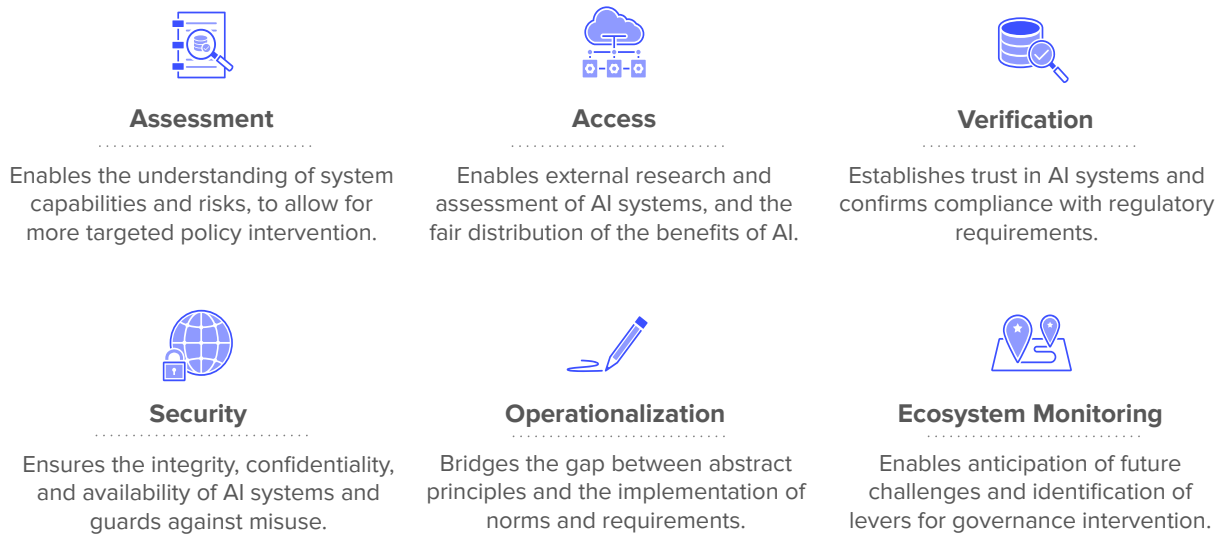
A Appendix: Policy Brief

The increasing adoption of artificial intelligence (AI) has prompted governance actions from the public sector, academia, civil society, and industry. However, policymakers often have insufficient information for identifying the need for intervention and assessing the efficacy of different policy options. Furthermore, the technical tools necessary for successfully implementing policy proposals are often lacking. We introduce the field of **technical AI governance**, which seeks to address these challenges.

Technical AI governance refers to *technical analysis and tools for supporting the effective governance of AI*. We argue that technical AI governance can:

1. ***Identify* areas where policy intervention is needed** through mapping technical aspects of systems to risks and opportunities associated with their application;
2. ***Inform* policy decisions** by assessing the effectiveness and feasibility of different policy options; and
3. ***Enhance* policy options** by enabling mechanisms for enforcing, incentivizing, or complying with norms and requirements.

We taxonomize technical AI governance according to elements of the AI value chain: the inputs of **data**, **compute**, **models**, and **algorithms**, through to the **deployment** setting of the resulting systems. The figure below shows the key governance capacities that can be applied to each target.



Key Takeaways

We highlight a number of the key takeaways within technical AI governance, including that:

- **Evaluations** of systems and their downstream impacts on users and society have been proposed in many governance regimes. However, current evaluations lack robustness, reliability, and validity, especially for foundation models.
- **Hardware mechanisms** could potentially enable actions including facilitating privacy-preserving access to datasets and models, verifying the use of computational resources, or attesting to the results of audits and evaluations. However, the use of such mechanisms for these purposes is largely unproven.

- The **development of infrastructure for enabling research into AI**, such as resources for conducting analyses of large training datasets or for providing privacy-preserving access to models for evaluation and auditing, could facilitate research that advances the scientific understanding of AI systems and external oversight into developers' activities.
- Research that aims to **monitor the AI ecosystem** by collecting and analyzing data on trends and advances in AI has already proven crucial for providing policymakers with the information needed to ensure that policy is forward-looking and future-proof.

We note that technical AI governance is merely one component of a comprehensive AI governance portfolio, and should be seen in service of sociotechnical and political solutions. A technosolutionist approach to AI governance and policy is unlikely to succeed.

Recommendations

Based on the above takeaways, we recommend:

1. **Allocating funding and resources** through open calls and funding bodies, to technical AI governance research, drawing on established expertise in adjacent fields;
2. That policymakers **collaborate closely with technical experts** to define feasible objectives and identify viable pathways to implementation;
3. That government bodies, such as AI Safety Institutes, **conduct in-house research on technical AI governance** topics, beyond their current focus on performing evaluations; and
4. That the future summits on AI, other fora such as the G7, the UN AI advisory body, and reports such as the *International Scientific Report on the Safety of Advanced AI*, **focus effort and attention towards technical AI governance**.

B Appendix: Methodology

We grounded our methodology in established practices from governance and policy research. Taxonomies are widely used in these fields to organize complex domains into structured categories (Smith, 2002). A taxonomy is essentially a "formal system for classifying multifaceted, complex phenomena according to common conceptual domains and dimensions" (Bradley et al. (2007), based on Patton (2002)). Typologies and taxonomies can help bring clarity to ill-defined problem spaces (Sofaer, 1999; Smith, 2002). Following this previous literature and work, we adopted a taxonomy-based approach to systematically map the landscape of technical AI governance challenges.

As outlined in Section 1, previous AI governance research, in particular, has introduced such taxonomies and frameworks to categorize governance issues. For example, Dafoe (2018) divided the AI governance space into three clusters – technical landscape, AI politics, and ideal governance structures – to attempt comprehensive coverage of AI governance challenges. Likewise, Critch & Russell (2023) proposed a taxonomy of societal-scale AI risks organized by accountability (i.e., who the responsible actors are and whether harmful actions are deliberate or emergent). These prior efforts illustrate that there is no single 'correct' AI governance taxonomy; rather, their value lies in choosing organizing principles that reveal new insights or gaps (Bailey, 1994).

For our research, we began with a systematic search on Google Scholar and SSRN using the search terms *'Technical' AND ('Artificial Intelligence' OR 'Machine Learning' OR 'AI' OR 'ML') AND ('Governance' OR 'Regulation')* to identify academic literature targeting technical AI governance, which did not yield any search results based on paper titles or abstracts. We then expanded search efforts to include the terms *('Artificial Intelligence' OR 'Machine Learning' OR 'AI' OR 'ML') AND ('Governance' OR 'Regulation')* to identify goals, functions, and challenges of AI governance more broadly, not just focused on technical work. In addition, we built on efforts by Reuel et al. (2024b) and compared existing regulatory and governance aims and the technical state of the art to identify gaps and governance challenges that could be addressed with technical advancements (these insights were later also integrated into the 'Motivation' sections for each open problem area in Sections 3 to 8).

Following this initial literature review, we generated a list of candidate capacities and open problems. We red-teamed this preliminary taxonomy with a focus group of five experts across AI and governance. Based on their feedback, which also referenced the work by Bommasani et al. (2023b), we added the *target* dimension to the taxonomy. We applied this mixture between deductive and inductive approaches throughout the whole taxonomy development process, in line with best practices for taxonomy development (Bailey, 1994; Nickerson et al., 2013). Our taxonomy then went through multiple rounds of expert- and literature-driven refinements. We iteratively discussed and adjusted capacity definitions in a series of interviews and focus groups with the authors, who are all researchers or practitioners in (AI) governance or relevant technical fields, and external experts, involving over 30 researchers and practitioners across all capacities, targets, and identified open problems areas. In these sessions, we critically examined whether each proposed category was conceptually distinct and collectively exhaustive of the domain's key issues and whether we missed any open problems. The resulting feedback led to taxonomy adjustments (e.g., clarifying the scope and definition of the *Operationalization* capacity) until we converged on a final taxonomy structure. Whenever a new capacity was added (including for the initial set of defined capacities), we also conducted a literature search on Google Scholar and SSRN with the terms *('Artificial Intelligence' OR 'AI') AND '[Capacity]'* to validate it, identify existing research in this area, and to expand on open problems in the identified area. We then again requested expert feedback on gaps, previous literature, open research questions, and missing open problems for the newly-added capacity.

We validated our final taxonomy's relevance by comparing it with previous classifications of issues in AI governance research and AI risk taxonomies, looking for overlaps and differences. We compared our open-problems categories to challenges raised in Dafoe (2018); Lobel (2024); Shelby et al. (2023); Raji et al. (2022a); Google (n.d.); Critch & Russell (2023); Weidinger et al. (2022); Birkstedt et al. (2023), confirming that all major governance challenge areas identified by these authors that could have a technical component have a place within our two-dimensional framework. In fact, many open problems we identify (such as evaluating

model risks or securing compute infrastructure) are technical prerequisites to addressing the higher-level governance challenges described in these prior works. For example, the challenge of governance failure due to inadequate oversight (Uuk et al., 2024) can be partially addressed by our verification capacity, and the risk of AI misuse (Critch & Russell, 2023) could be mitigated by advances in security and assessment, two of our capacities which are also governance mechanisms highlighted by McCormack & Bendechache (2024) and Lobel (2024). This cross-validation across capacities gave us confidence that our taxonomy is neither missing well-known challenge domains nor duplicating existing taxonomies without added value. Instead, it synthesizes and builds on prior efforts: for instance, where Critch & Russell (2023) focuses on *who* is responsible for risks, and Shelby et al. (2023) catalog harms by *impact type*, our taxonomy focuses on *how* technical work can proactively support governance.

Finally, we acknowledge our taxonomy's limitations. It is expert-defined and thus subject to unique perspectives and experiences. Furthermore, validation is challenging, given that there is no ground truth for our taxonomy to test against. It is further constrained to technical challenge areas by design, meaning it leaves out many important socio-technical or institutional governance issues (e.g. labor impacts, international dynamics) highlighted by others (Dafoe, 2018; Lobel, 2024; Shelby et al., 2023; Raji et al., 2022a; Critch & Russell, 2023; Weidinger et al., 2022).

References

Onni Aarne, Tim Fist, and Caleb Withers. Secure, governable chips: Using On-Chip mechanisms to manage national security risks from AI & advanced computing. Technical report, Center for a New American Security, 2024. URL https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf.

Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei, and Laura Waltersdorfer. A collaborative, Human-Centred taxonomy of AI, algorithmic, and automation harms. *arXiv: 2407.01294 [cs.LG]*, July 2024. URL http://arxiv.org/abs/2407.01294.

Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51(4):1–35, July 2018. ISSN 0360-0300. URL https://doi.org/10.1145/3214303.

Alessandro Achille, Michael Kearns, Carson Klingenberg, and Stefano Soatto. AI model disgorgement: Methods and choices. *Proceedings of the National Academy of Sciences of the United States of America*, 121(18):e2307304121, April 2024. ISSN 0027-8424, 1091-6490. URL http://dx.doi.org/10.1073/pnas.2307304121.

Fran Ackermann, Susan Howick, John Quigley, Lesley Walls, and Tom Houghton. Systemic risk elicitation: Using causal maps to engage stakeholders and build a comprehensive view of risks. *European journal of operational research*, 238(1):290–299, October 2014. ISSN 0377-2217. URL https://www.sciencedirect.com/science/article/pii/S0377221714002744.

Ada Lovelace Institute. Keeping an eye on AI: Approaches to government monitoring of the AI landscape. Technical report, Ada Lovelace Institute, July 2023. URL https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/.

Advisory Body on Artificial Intelligence. Interim report: Governing AI for humanity. Technical report, United Nations, December 2023. URL https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf.

Nur Ahmed and Muntasir Wahed. The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv: 2010.15581 [cs.CY]*, October 2020. doi: 10.48550/arXiv.2010.15581.

Mhairi Aitken, David Leslie, Florian Ostmann, Jacob Pratt, Helen Margetts, and Cosmina Dorobantu. Common regulatory capacity for AI. Technical report, The Alan Turing Institute, 2022. URL https://www.turing.ac.uk/news/publications/common-regulatory-capacity-ai.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv: 2305.16867 [cs.CL]*, May 2023. URL http://arxiv.org/abs/2305.16867.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. Towards tracing factual knowledge in language models back to the training data. *arXiv: 2205.11482 [cs.CL]*, May 2022. URL http://arxiv.org/abs/2205.11482.

John Albert. A guide to the EU's new rules for researcher access to platform data, 2022. URL https://algorithmwatch.org/en/dsa-data-access-explained/. Accessed: 2024-7-17.

Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Déforges. Adversarial example detection for DNN models: a review and experimental comparison. *Artificial intelligence review*, 55(6): 4403–4462, August 2022. ISSN 0269-2821, 1573-7462. URL https://doi.org/10.1007/s10462-021-10125-w.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-Consuming generative models go MAD. In *The 12th International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, October 2023. URL https://openreview.net/forum?id=ShjMHfmPs0.

Gregory C Allen. Choking off china's access to the future of AI. Technical report, Center for Strategic and International Studies, 2022. URL https://www.csis.org/analysis/choking-chinas-access-future-ai.

Gregory C Allen, Emily Benson, and William Alan Reinsch. Improved export controls enforcement technology needed for U.S. national security. Technical report, Center for Strategic and International Studies (CSIS), 2022. URL http://www.jstor.org/stable/resrep53648.

Hilary J Allen. Fintech and Techno-Solutionism, January 2024. URL https://papers.ssrn.com/abstract=4686469.

AMD. AMD secure encrypted virtualization (SEV), 2024. URL https://www.amd.com/en/developer/sev.html. Accessed: 2024-7-17.

Markus Anderljung, Lennart Heim, and Toby Shevlane. Compute funds and pre-trained models, April 2022. URL https://www.governance.ai/post/compute-funds-and-pre-trained-models. Accessed: 2022-7-3.

Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier AI regulation: Managing emerging risks to public safety. *arXiv: 2307.03718 [cs.CY]*, July 2023a. doi: 10.48550/arXiv.2307.03718.

Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework. *arXiv: 2311.14711 [cs.CY]*, November 2023b. URL http://arxiv.org/abs/2311.14711.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading Safety-Aligned LLMs with simple adaptive attacks. *arXiv: 2404.02151 [cs.CR]*, April 2024. URL http://arxiv.org/abs/2404.02151.

María P Angel and Danah Boyd. Techno-legal solutionism: Regulating children's online safety in the united states. In *Proceedings of the Symposium on Computer Science and Law*, CSLAW '24, pp. 86–97, New York, NY, USA, March 2024. Association for Computing Machinery. ISBN 9798400703331. URL https://doi.org/10.1145/3614407.3643705.

Anthropic. Anthropic's responsible scaling policy, September 2023a. URL https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed: 2024-5-6.

Anthropic. Challenges in evaluating AI systems, 2023b. URL https://www.anthropic.com/news/evaluating-ai-systems. Accessed: 2024-7-16.

Anthropic. Introducing the next generation of claude, March 2024a. URL https://www.anthropic.com/news/claude-3-family. Accessed: 2024-4-17.

Anthropic. Claude 3.5 sonnet, 2024b. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-7-15.

Emily Apsey, Phil Rogers, Michael O'Connor, and Rob Nertney. Confidential computing on NVIDIA H100 GPUs for secure and trustworthy AI, August 2023. URL https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/. Accessed: 2024-7-17.

Rudolf Avenhaus, Nicholas Kyriakopoulos, Michel Richard, and Gotthard Stein (eds.). *Verifying Treaty Compliance*. Springer Berlin Heidelberg, 2006. URL https://link.springer.com/book/10.1007/3-540-33854-3.

Kenneth D Bailey. *Typologies and taxonomies: An introduction to classification techniques*, volume 12. Sage, 1994.

Mauricio Baker. Nuclear arms control verification and lessons for AI treaties. *arXiv: 2304.04123 [cs.CY]*, April 2023. URL http://arxiv.org/abs/2304.04123.

Shyamkrishna Balganesh. The normativity of copying in copyright law. *Duke law journal*, 2012. ISSN 0012-7086. URL https://scholarship.law.upenn.edu/faculty_scholarship/702.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156. IEEE, May 2022. ISBN 9781665413169, 9781665413176. URL http://dx.doi.org/10.1109/SP46214.2022.9833677.

Bank for International Settlements. BIS innovation hub expands suptech and regtech research to include monetary policy tech, February 2021. URL https://www.bis.org/about/bisih/topics/suptech_regtech.htm. Accessed: 2024-7-12.

Peter Barnett, Rachel Freedman, Justin Svegliato, and Stuart Russell. Active reward learning from multiple teachers. *arXiv: 2303.00894 [cs.LG]*, March 2023. URL http://arxiv.org/abs/2303.00894.

Anthony Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R Murphy, and Krystal Jackson. AI Risk-Management standards profile for General-Purpose AI systems (GPAIS) and foundation models. Technical report, UC Berkeley Center for Long-term Cybersecurity, 2023. URL https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/.

Adrien Basdevant, Camille François, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, Nathan Lambert, Stefano Maffulli, Nik Marda, Govind Shivkumar, and Justine Tunney. Towards a framework for openness in foundation models: Proceedings from the columbia convening on openness in artificial intelligence. *arXiv: 2405.15802 [cs.SE]*, May 2024. URL http://arxiv.org/abs/2405.15802.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv: 2006.14651 [cs.LG]*, June 2020. URL http://arxiv.org/abs/2006.14651.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=awIpKpwTwF&noteId=Ju4XcafMir.

Yohsua Bengio, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, Vasilios Mavroudis, Mantas Mazeika, Kwan Yee Ng, Chinasa T Okolo, Deborah Raji, Theodora Skeadas, Florian Tramèr, and Soren Mindermann. International scientific report on the safety of advanced AI. Technical report, Department for Science, Innovation and Technology, 2024. URL https://hal.science/hal-04612963/.

Bennett Institute for Applied Data Science. About OpenSAFELY, 2024. URL https://www.opensafely.org/about/. Accessed: 2024-7-17.

Tony Berber Sardinha. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083, April 2024. ISSN 2666-7991. URL https://www.sciencedirect.com/science/article/pii/S2666799123000436.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety – a review. *arXiv: 2404.14082 [cs.AI]*, April 2024. URL http://arxiv.org/abs/2404.14082.

Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal adaptation to advanced AI. *arXiv: 2405.10295 [cs.CY]*, May 2024. URL http://arxiv.org/abs/2405.10295.

Tamay Besiroglu, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv: 2401.02452 [cs.CY]*, January 2024. doi: 10.48550/arXiv.2401.02452.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv: 2405.14782 [cs.CL]*, May 2024. URL http://arxiv.org/abs/2405.14782.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1467–1474, Madison, WI, USA, June 2012. Omnipress. ISBN 9781450312851. URL https://dl.acm.org/doi/10.5555/3042573.3042761.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv: 2110.01963 [cs.CY]*, October 2021. URL http://arxiv.org/abs/2110.01963.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. Into the LAIONs den: Investigating hate in multimodal datasets. *arXiv: 2311.03449 [cs.CY]*, November 2023. URL http://arxiv.org/abs/2311.03449.

Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. Ai governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

Cody Blakeney, Mansheej Paul, Brett W Larsen, Sean Owen, and Jonathan Frankle. Does your data spark joy? Performance gains from domain upsampling at the end of training. *arXiv: 2406.03476 [cs.LG]*, June 2024. URL http://arxiv.org/abs/2406.03476.

Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 5454–5476, Virtual, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.485.

Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. Exploring the relevance of data Privacy-Enhancing technologies for AI governance use cases. *arXiv: 2303.08956 [cs.AI]*, March 2023. URL http://arxiv.org/abs/2303.08956.

Miranda Bogen and Amy Winecoff. Applying sociotechnical approaches to AI governance in practice. Technical report, Center for Democracy & Technology, May 2024. URL https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice/.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. Technical report, Center for Research on Foundation Models (CRFM) and Institute on Human-Centered Artificial Intelligence (HAI), October 2023a. URL http://arxiv.org/abs/2310.12941.

Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang. Do foundation model providers comply with the draft EU AI act? Technical report, Stanford Center for Research on Foundation Models, 2023b. URL https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *arXiv: 2303.15772 [cs.LG]*, March 2023c. doi: 10.48550/arXiv.2303.15772.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. *arXiv: 2402.16268 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.16268.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, Virtual, May 2021. IEEE. ISBN 9781728189345, 9781728189352. URL http://dx.doi.org/10.1109/SP40001.2021.00019.

Samuel R Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.385.

Elizabeth H Bradley, Leslie A Curry, and Kelly J Devers. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health services research*, 42(4):1758–1772, 2007.

Asher Brass and Onni Aarne. Location verification for AI chips. Technical report, Institute for AI Policy and Strategy, 2024. URL https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6670467ebe2a477eb1554f40/1718634112482/Location%2BVerification%2Bfor%2BAI%2BChips.pdf.

Paul Bricman. Hashmarks: Privacy-Preserving benchmarks for High-Stakes AI evaluation. *arXiv: 2312.00645 [cs.LG]*, December 2023. URL http://arxiv.org/abs/2312.00645.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 2280–2292, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://dl.acm.org/doi/10.1145/3531146.3534642.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J B Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv: 2004.07213 [cs.CY]*, April 2020. URL http://arxiv.org/abs/2004.07213.

Benjamin S Bucknall and Robert F Trager. Structured access for third-party research on frontier ai models: Investigating researchers' model access requirements. Technical report, Oxford Martin School, University of Oxford and Center for the Governance of AI, October 2023. URL https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf.

Justin B Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M Hudson, Anton Korinek, Matthew M Young, and Baobao Zhang (eds.). *The Oxford Handbook of AI Governance*. Oxford University Press, London, England, 1 edition, February 2022. ISBN 9780197579329, 9780197579350. URL https://academic.oup.com/edited-volume/41989.

Bureau of Industry and Security. Commerce implements new export controls on advanced computing and semiconductor manufacturing items to the people's republic of china (PRC). Technical report, U.S. Department of Commerce, 2022a.

Bureau of Industry and Security. Implementation of additional export controls: Certain advanced computing and semiconductor manufacturing items; supercomputer and semiconductor end use; entity list modification, October 2022b. URL https://www.federalregister.gov/d/2022-21658.

Bureau of Industry and Security. Implementation of additional export controls: Certain advanced computing items; supercomputer and semiconductor end use; updates and corrections, October 2023. URL https://www.federalregister.gov/d/2023-23055.

Bureau of Industry and Security. Commerce control List—Supplement no. 1 to part 774—category 4, 2024. URL https://www.bis.doc.gov/index.php/documents/regulations-docs/2335-ccl4-5/file.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv: 2212.03827 [cs.CL]*, December 2022. doi: 10.48550/arXiv.2212.03827.

C2PA. Overview, 2022. URL https://c2pa.org/. Accessed: 2024-7-17.

Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image, and video technology*, 41(3):271–284, November 2005. ISSN 1387-5485, 0922-5773. URL https://doi.org/10.1007/s11265-005-4151-3.

Nicholas Carlini. Poisoning the unlabeled dataset of Semi-Supervised learning. *arXiv: 2105.01622 [cs.LG]*, May 2021. URL https://arxiv.org/abs/2105.01622.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv: 2106.09667 [cs.LG]*, June 2021. URL http://arxiv.org/abs/2106.09667.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 9781939133069. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models. In Daniele Micciancio and Thomas Ristenpart (eds.), *Advances in Cryptology – CRYPTO 2020*, pp. 189–218, Cham, 2020. Springer International Publishing. ISBN 9783030568771. URL http://dx.doi.org/10.1007/978-3-030-56877-1_7.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX security symposium (USENIX security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 9781939133243. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13263–13276. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/564b5f8289ba846ebc498417e834c253-Paper-Conference.pdf.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning Web-Scale training datasets is practical. *arXiv: 2302.10149 [cs.CR]*, February 2023a. URL http://arxiv.org/abs/2302.10149.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, November 2023b. URL https://openreview.net/forum?id=OQQoD8Vc3B.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. *arXiv: 2403.06634 [cs.CR]*, March 2024. doi: 10.48550/arXiv.2403.06634.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX security 23)*, pp. 5253–5270, Anaheim, CA, August 2023c. USENIX Association. ISBN 9781939133373. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. Deceptive alignment monitoring. *arXiv: 2307.10569 [cs.LG]*, July 2023. URL http://arxiv.org/abs/2307.10569.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv: 2306.09442 [cs.CL]*, June 2023. doi: 10.48550/arXiv.2306.09442.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box access is insufficient for rigorous AI audits. *arXiv: 2401.14446 [cs.CY]*, January 2024a. URL http://arxiv.org/abs/2401.14446.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv: 2403.05030 [cs.CR]*, March 2024b. doi: 10.48550/arXiv.2403.05030.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7):1–38, April 2024. ISSN 0360-0300. URL https://doi.org/10.1145/3616865.

Iker Ceballos, Vivek Sharma, Eduardo Mugica, Abhishek Singh, Alberto Roman, Praneeth Vepakomma, and Ramesh Raskar. SplitNN-driven vertical partitioning. *arXiv: 2008.04137 [cs.LG]*, August 2020. URL http://arxiv.org/abs/2008.04137.

Alan Chan. Evaluating predictions of model behaviour, 2024. URL https://www.governance.ai/post/evaluating-predictions-of-model-behaviour. Accessed: 2024-7-16.

Alan Chan, Chinasa T Okolo, Zachary Terner, and Angelina Wang. The limits of global inclusion in AI development. *arXiv: 2102.01265 [cs.CY]*, February 2021. URL http://arxiv.org/abs/2102.01265.

Alan Chan, Maxime Riché, and Jesse Clifton. Towards the scalable evaluation of cooperativeness in language models. *arXiv: 2303.13360 [cs.CL]*, March 2023a. URL http://arxiv.org/abs/2303.13360.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 651–666, New York, NY, USA, June 2023b. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI agents. *arXiv: 2401.13138 [cs.CY]*, January 2024. doi: 10.48550/arXiv.2401.13138.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):39:1–39:45, March 2024. ISSN 2157-6904, 2157-6912. URL https://doi.org/10.1145/3641289.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv: 2404.01318 [cs.CR]*, March 2024. URL http://arxiv.org/abs/2404.01318.

Brian J Chen and Jacob Metcalf. Explainer: A sociotechnical approach to AI policy. Technical report, Data & Society, 2024. URL https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/.

Huili Chen, Cheng Fu, Bita Darvish Rouhani, Jishen Zhao, and Farinaz Koushanfar. DeepAttest: an end-to-end attestation framework for deep neural networks. In *Proceedings of the 46th International Symposium on Computer Architecture*, ISCA '19, pp. 487–498, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 9781450366694. URL https://doi.org/10.1145/3307650.3322251.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, Ion Stoica, and Eric P Xing. Vicuna: An Open-Source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Hyeongmin Cho and Sangkyun Lee. Data quality measures and efficient evaluation algorithms for Large-Scale High-Dimensional data. *arXiv: 2101.01441 [cs.LG]*, January 2021. URL http://arxiv.org/abs/2101.01441.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. What is your data worth to GPT? LLM-Scale data valuation with influence functions. *arXiv: 2405.13954 [cs.LG]*, May 2024. URL http://arxiv.org/abs/2405.13954.

Dami Choi, Yonadav Shavit, and David K Duvenaud. Tools for verifying neural models' training data. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track*, volume abs/2307.00682. 2024, July 2023. URL http://dx.doi.org/10.48550/arXiv.2307.00682.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, June 2017. ISSN 2167-647X, 2167-6461. URL http://dx.doi.org/10.1089/big.2016.0047.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv: 2306.09194 [cs.CR]*, May 2023. doi: 10.48550/arXiv.2306.09194.

Jack Clark. Information markets and AI development. In *The Oxford Handbook of AI Governance*, pp. 345–357. Oxford University Press, January 2023. ISBN 9780197579329, 9780197579350. URL https://academic.oup.com/edited-volume/41989/chapter/393374951.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1310–1320. PMLR, May 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Michael K Cohen, Noam Kolt, Yoshua Bengio, Gillian K Hadfield, and Stuart Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, April 2024a. URL https://www.science.org/doi/10.1126/science.adl0625.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, April 2024b. ISSN 2307-387X. URL https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00644/2362212/tacl_a_00644.pdf.

Commerce Department. Taking additional steps to address the national emergency with respect to significant malicious Cyber-Enabled activities, 01 2024. URL https://www.federalregister.gov/d/2024-01580.

Content Authenticity Initiative. Restoring trust and transparency in the age of AI, 2024. URL https://contentauthenticity.org/. Accessed: 2024-7-17.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023. URL https://ieeexplore.ieee.org/document/10095167.

Council of the European Union. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act), 2024. URL https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf.

Andrew Critch and Stuart Russell. TASRA: A taxonomy and analysis of Societal-Scale risks from AI. *arXiv: 2306.06924 [cs.AI]*, June 2023. URL http://arxiv.org/abs/2306.06924.

Rachel Crowell. Why AI's diversity crisis matters, and how to tackle it. *Nature*, May 2023. ISSN 0028-0836, 1476-4687. URL https://www.nature.com/articles/d41586-023-01689-4.

Cybersecurity and Infrastructure Security Agency. Secure by design, 2023. URL https://www.cisa.gov/securebydesign. Accessed: 2024-7-18.

Allan Dafoe. AI governance: A research agenda. Technical report, Centre for the Governance of AI & Future of Humanity Institute, University of Oxford, 2018. URL http://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

Morten Dahl. Private deep learning with MPC: A simple tutorial from scratch, 2017. URL https://mortendahl.github.io/2017/04/17/private-deep-learning-with-mpc/. Accessed: 2024-7-17.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv: 2405.06624 [cs.AI]*, May 2024. URL http://arxiv.org/abs/2405.06624.

Angela Daly, Thilo Hagendorff, Li Hui, Monique Mann, Vidushi Marda, Ben Wagner, Wayne Wei Wang, Hans-W Micklitz, Oreste Pollicino, Amnon Reichman, Andrea Simoncini, Giovanni Sartor, and Giovanni De Gregorio. AI, governance and ethics. In *Constitutional Challenges in the Algorithmic Society*, pp. 182–201. Cambridge University Press, December 2021. URL https://www.cambridge.org/core/product/C3C08005487663E5BE66FF72690DC8FA.

Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. AI capabilities can be significantly improved without expensive retraining. Technical report, Epoch, December 2023. URL http://arxiv.org/abs/2312.07413.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv: 2107.07002 [cs.LG]*, July 2021. URL http://arxiv.org/abs/2107.07002.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*, pp. 1–23, New York, NY, USA, October 2023. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3617694.3623261.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. *arXiv: 2311.09783 [cs.CL]*, November 2023. URL http://arxiv.org/abs/2311.09783.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. SOPHON: Non-Fine-Tunable learning to restrain task transferability for pre-trained models. *arXiv: 2404.12699 [cs.LG]*, April 2024. URL http://arxiv.org/abs/2404.12699.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. *arXiv: 2205.12548 [cs.CL]*, May 2022. URL http://arxiv.org/abs/2205.12548.

Department for Science, Innovation & Technology. Emerging processes for frontier AI safety. Technical report, GOV.UK, October 2023. URL https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf.

Department for Science, Innovation & Technology and AI Safety Institute. Advanced AI evaluations at AISI: May update. Technical report, GOV.UK, 2024. URL https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

Department for Science, Innovation and Technology and Office for Artificial Intelligence. A pro-innovation approach to AI regulation. Technical report, GOV.UK, August 2023. URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

Department for Science, Innovation and Technology, Foreign, Commonwealth & Development Office, and Prime Minister's Office, 10 Downing Street. The bletchley declaration by countries attending the AI safety summit, 1-2 november 2023. Technical report, GOV.UK, November 2023a. URL https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Department for Science, Innovation and Technology, Foreign, Commonwealth & Development Office, and Prime Minister's Office 10 Downing Street. Chair's summary of the AI safety summit 2023, bletchley park. Technical report, GOV.UK, 2023b. URL https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park.

Gobikrishna Dhanuskodi, Sudeshna Guha, Vidhya Krishnan, Aruna Manjunatha, Michael O'Connor, Rob Nertney, and Phil Rogers. Creating the first confidential GPUs: The team at NVIDIA brings confidentiality and integrity to user code and data for accelerated computing. *Queueing Systems. Theory and Applications*, 21(4):68–93, September 2023. ISSN 0257-0130, 1542-7730. URL https://doi.org/10.1145/3623393.3623391.

Digital Services Act. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance), October 2022. URL http://data.europa.eu/eli/reg/2022/2065/oj/eng.

Roel Dobbe and Anouk Wolters. Toward sociotechnical AI: Mapping vulnerabilities for machine learning in context. *Minds and Machines*, 34(2):12, May 2024. ISSN 0924-6495, 1572-8641. URL https://doi.org/10.1007/s11023-024-09668-y.

Roel I J Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments (AIES '20). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 242, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 9781450371100. URL https://doi.org/10.1145/3375627.3375861.

Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. A sociotechnical view of algorithmic fairness. *Information systems journal*, 32(4):754–818, 2022. ISSN 1350-1917, 1365-2575. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/isj.12370.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding large language models: A survey. *arXiv: 2406.02622 [cs.CR]*, June 2024a. URL http://arxiv.org/abs/2406.02622.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv: 2402.15938 [cs.CL]*, February 2024b. URL http://arxiv.org/abs/2402.15938.

Arthur Douillard, Qixuan Feng, Andrei A Rusu, Adhiguna Kuncoro, Yani Donchev, Rachita Chhaparia, Ionel Gog, Marc'aurelio Ranzato, Jiajun Shen, and Arthur Szlam. DiPaCo: Distributed path composition. *arXiv: 2403.10616 [cs.LG]*, March 2024. URL http://arxiv.org/abs/2403.10616.

Anca Dragan, Helen King, and Allan Dafoe. Introducing the frontier safety framework, 2024. URL https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/. Accessed: 2024-7-12.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv: 2402.07841 [cs.CL]*, February 2024a. URL http://arxiv.org/abs/2402.07841.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. Uncovering latent memories: Assessing data leakage and memorization patterns in large language models. *arXiv: 2406.14549 [cs.CV]*, June 2024b. URL http://arxiv.org/abs/2406.14549.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent AI: Surveying the horizons of multimodal interaction. *arXiv: 2401.03568 [cs.AI]*, January 2024. URL http://arxiv.org/abs/2401.03568.

Dynabench. Challenging the limits of benchmarking AI, 2023. URL https://dynabench.org/. Accessed: 2024-7-16.

Colin Eden, Fran Ackermann, Charles B Finn, and John M Bryson. *Visible Thinking: Unlocking Causal Mapping for Practical Business Results*. John Wiley & Sons, 2004. ISBN 9780470869161. URL https://play.google.com/store/books/details?id=LLjEuBlSUoMC.

Janet Egan and Lennart Heim. Oversight for frontier AI through a Know-Your-Customer scheme for compute providers. *arXiv: 2310.13625 [cs.CY]*, October 2023. URL http://arxiv.org/abs/2310.13625.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.

Ronen Eldan and Mark Russinovich. Who's harry potter? Approximate unlearning in LLMs. *arXiv: 2310.02238 [cs.CL]*, October 2023. doi: 10.48550/arXiv.2310.02238.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, June 2024. ISSN 0036-8075, 1095-9203. URL http://dx.doi.org/10.1126/science.adj0998.

Epoch. Key trends and figures in machine learning, 2023. URL https://epochai.org/trends.

EU Joint Research Centre. FAQs: DSA data access for researchers, 2023. URL https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en. Accessed: 2024-7-17.

European Commission. EU general data protection regulation (GDPR): Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). *OJ 2016 L 119/1*, 59:1–88, May 2016.

European Commission. Commission welcomes G7 leaders' agreement on guiding principles and a code of conduct on artificial intelligence, October 2023. URL https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-g7-leaders-agreement-guiding-principles-and-code-conduct-artificial. Accessed: 2024-7-12.

David Evans, Vladimir Kolesnikov, and Mike Rosulek. *A Pragmatic Introduction to Secure Multi-Party Computation*. Now Foundations and Trends, 2018. ISBN 9781680835083, 9781680835090. URL http://dx.doi.org/10.1561/3300000019.

Executive Office of the President. Safe, secure, and trustworthy development and use of artificial intelligence, November 2023. URL https://www.federalregister.gov/d/2023-24283.

Congyu Fang, Hengrui Jia, Anvith Thudi, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-Learning is currently more broken than you think. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 797–816. IEEE, July 2023. ISBN 9781665465120, 9781665465137. URL http://dx.doi.org/10.1109/EuroSP57164.2023.00052.

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. LLM agents can autonomously hack websites. *arXiv: 2402.06664 [cs.CR]*, February 2024. doi: 10.48550/arXiv.2402.06664.

Uriel Fiege, Amos Fiat, and Adi Shamir. Zero knowledge proofs of identity. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pp. 210–217, 1987.

Tim Fist and Erich Grunewald. Preventing AI chip smuggling to china. Technical report, Center for a New American Security, October 2023. URL https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china.

Jordan Frery, Andrei Stoian, Roman Bredehoft, Luis Montero, Celia Kherfallah, Benoit Chevallier-Mames, and Arthur Meyre. Privacy-Preserving Tree-Based inference with fully homomorphic encryption, 2023. URL https://eprint.iacr.org/2023/258.

Meir Friedenberg and Joseph Y Halpern. Blameworthiness in Multi-Agent settings. *arXiv: 1903.04102 [cs.CY]*, March 2019. URL http://arxiv.org/abs/1903.04102.

Frontier Model Forum. Issue brief: Measuring training compute, 2024. URL https://www.frontiermodelforum.org/updates/issue-brief-measuring-training-compute/. Accessed: 2024-7-18.

G7. Hiroshima process international guiding principles for organizations developing advanced ai system. 2023. URL https://www.mofa.go.jp/files/100573471.pdf.

G7 leaders. Hiroshima process international code of conduct for advanced AI systems, 2023. URL https://www.mofa.go.jp/files/100573473.pdf.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced AI assistants. Technical report, Google DeepMind, April 2024. URL http://arxiv.org/abs/2404.16244.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. Technical report, Anthropic, November 2022. URL http://arxiv.org/abs/2209.07858.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, Honolulu, Hawaii, USA, July 2023a. PMLR. URL https://proceedings.mlr.press/v202/gao23h.html.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. lm-evaluation-harness: A framework for few-shot language model evaluation, 2023b. URL https://github.com/EleutherAI/lm-evaluation-harness.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv: 2406.04093 [cs.LG]*, June 2024. URL http://arxiv.org/abs/2406.04093.

Yansong Gao, Said F Al-Sarawi, and Derek Abbott. Physical unclonable functions. *Nature Electronics*, 3(2): 81–91, February 2020. ISSN 2520-1131, 2520-1131. URL https://www.nature.com/articles/s41928-020-0372-5.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, November 2021. ISSN 0001-0782. URL https://dl.acm.org/doi/10.1145/3458723.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, Yaguang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, Hyunjeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi

Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M R Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-Yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Li, Music, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern

Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, M, Pavan Kumar Reddy, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z J Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi Lv, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, R, Raghavender, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer

Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian Lin, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, Mohammadhossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser Tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A Choquette-Choo, Yunjie Li, T J Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, Xianghai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M K Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua

Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models. Technical report, Google DeepMind, 2023. URL http://arxiv.org/abs/2312.11805.

Craig Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, Stanford, CA, USA, 2009. URL https://dl.acm.org/doi/10.5555/1834954.

Tim Geppert, Stefan Deml, David Sturzenegger, and Nico Ebert. Trusted execution environments: Applications and organizational challenges. *Frontiers in Computer Science*, 4, 2022. ISSN 2624-9898. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.930741.

Tom Gerken and Imran Rahman-Jones. Rishi sunak: AI firms cannot 'mark their own homework'. *BBC*, November 2023. URL https://www.bbc.com/news/technology-67285315.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A Roberts, Diyi Yang, David L Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv: 2404.01413 [cs.LG]*, April 2024. URL http://arxiv.org/abs/2404.01413.

Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *8th USENIX symposium on networked systems design and implementation (NSDI 11)*, 2011. URL https://www.usenix.org/conference/nsdi11/dominant-resource-fairness-fair-allocation-multiple-resource-types.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251, New Orleans, LA, USA, 2019. PMLR. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Bedi. A survey on the possibilities & impossibilities of AI-generated text detection. *Transactions on Machine Learning Research*, October 2023. ISSN 2835-8856. URL https://openreview.net/pdf?id=AXtFeYjboj.

Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 84–130. Association for Computing Machinery, 2023.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. IEEE, October 2018. ISBN 9781538650905, 9781538650912. URL http://dx.doi.org/10.1109/DSAA.2018.00018.

David Glukhov, Ziwen Han, Ilia Shumailov, Vardan Papyan, and Nicolas Papernot. A false sense of safety: Unsafe information leakage in 'safe' AI responses. *arXiv: 2407.02551 [cs.CR]*, July 2024. URL http://arxiv.org/abs/2407.02551.

Augustin Godinot, Erwan Le Merrer, Gilles Trédan, Camilla Penzo, and Franois Taïani. Under manipulations, are some ai models harder to audit? In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 644–664, Los Alamitos, CA, USA, apr 2024. IEEE Computer Society. doi: 10.1109/SaTML59370.2024.00038. URL https://doi.ieeecomputersociety.org/10.1109/SaTML59370.2024.00038.

Shashwat Goel, Ameya Prabhu, Philip Torr, Ponnurangam Kumaraguru, and Amartya Sanyal. Corrective machine unlearning. *arXiv: 2402.14015 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.14015.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The 12th International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, October 2023. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Oded Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110):1–108, 1998.

Shafi Goldwasser, Guy N Rothblum, Jonathan Shafer, and Amir Yehudayoff. Interactive proofs for verifying machine learning. In James R Lee (ed.), *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 41:1–41:19, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 9783959771771. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2021.41.

Google. Perspectives on issues in ai governance. n.d. URL https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Google Cloud. Remote attestation of disaggregated machines, 2024. URL https://cloud.google.com/docs/security/remote-attestation. Accessed: 2024-7-17.

GOV.UK. AI safety institute: Overview, November 2023. URL https://www.gov.uk/government/publications/ai-safety-institute-overview. Accessed: 2024-7-16.

Declan Grabb, Max Lamparth, and Nina Vasan. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation, 2024. URL https://www.medrxiv.org/content/early/2024/04/08/2024.04.07.24305462.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. *arXiv: 2312.06942 [cs.LG]*, December 2023. doi: 10.48550/arXiv.2312.06942.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R Bowman. Studying large language model generalization with influence functions. *arXiv: 2308.03296 [cs.LG]*, August 2023. URL http://arxiv.org/abs/2308.03296.

Erich Grunewald and Michael Aird. AI chip smuggling into china: Potential paths, quantities, and countermeasures. Technical report, Institute for AI Policy and Strategy, 2023. URL https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651bb8a18f961e3333e3c1d7/1696315558319/AI+chip+smuggling+into+China+%5Bfinal%5D.pdf.

Neel Guha, Christie M Lawrence, Lindsey A Gailmard, Kit T Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, and Daniel E Ho. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *The George Washington Law Review*, 92(forthcoming), 2024. ISSN 0016-8076. URL https://dho.stanford.edu/wp-content/uploads/AI_Regulation.pdf.

M Guihot, A F Matthew, and N P Suzor. Nudging robots: Innovative solutions to regulate artificial intelligence. *Vand. J. Ent. & Tech. L.*, 2017. URL https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/vanep20&section=16.

Odd Erik Gundersen, Saeid Shamsaliei, and Richard Juul Isdahl. Do machine learning platforms provide out-of-the-box reproducibility? *Future generations computer systems: FGCS*, 126:34–47, January 2022. ISSN 0167-739X. URL https://www.sciencedirect.com/science/article/pii/S0167739X21002090.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. *arXiv: 2104.13733 [cs.CL]*, April 2021. URL http://arxiv.org/abs/2104.13733.

Chuan Guo, Awni Hannun, Brian Knott, Laurens van der Maaten, Mark Tygert, and Ruiyu Zhu. Secure multiparty computations in floating-point arithmetic. *Information and Inference*, 11(1):103–135, March 2022. URL https://academic.oup.com/imaiai/article-pdf/11/1/103/43152514/iaaa038.pdf.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv: 2311.09807 [cs.CL]*, November 2023. URL http://arxiv.org/abs/2311.09807.

Alexa Hagerty and Igor Rubinov. Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. *arXiv: 1907.07892 [cs.CY]*, July 2019. doi: 10.48550/arXiv.1907.07892.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. *arXiv: 2502.14143 [cs.MA]*, February 2025. URL http://arxiv.org/abs/2502.14143.

Hammond et al., forthcoming.

Krishnaprasad Hande. Announcing azure confidential VMs with NVIDIA H100 tensor core GPUs in preview, 2023. URL https://techcommunity.microsoft.com/t5/azure-confidential-computing/announcing-azure-confidential-vms-with-nvidia-h100-tensor-core/ba-p/3975389. Accessed: 2024-7-17.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in AI: Challenges, applications, and ethical implications. *arXiv: 2401.01629 [cs.LG]*, January 2024. URL http://arxiv.org/abs/2401.01629.

Barath Harithas. Mapping the chip smuggling pipeline and improving export control compliance. Technical report, Center for Strategic and International Studies, 2024. URL https://www.csis.org/analysis/mapping-chip-smuggling-pipeline-and-improving-export-control-compliance.

Lennart Heim. A trusted AI compute cluster for AI verification and evaluation, 2024. URL https://blog.heim.xyz/a-trusted-ai-compute-cluster/. Accessed: 2024-7-17.

Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael Osborne, and Noa Zilberman. Governing through the cloud: The intermediary role of compute providers in AI regulation. Technical report, Oxford Martin AI Governance Initiative, March 2024. URL https://cdn.governance.ai/Governing-Through-the-Cloud_The-Intermediary-Role-of-Compute-Providers-in-AI-Regulation.pdf.

William Held, Camille Harris, Michael Best, and Diyi Yang. A material lens on coloniality in NLP. *arXiv: 2311.08391 [cs.CL]*, November 2023. URL http://arxiv.org/abs/2311.08391.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *arXiv: 2303.15715 [cs.CY]*, March 2023a. URL http://arxiv.org/abs/2303.15715.

Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. Self-Destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 287–296, New York, NY, USA, August 2023b. Association for Computing Machinery. ISBN 9798400702310. URL https://doi.org/10.1145/3600211.3604690.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *The 9th International Conference on Learning Representations (ICLR 2021)*, Virtual, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv: 2306.12001 [cs.CY]*, June 2023. doi: 10.48550/arXiv.2306.12001.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv: 2304.00740 [cs.CL]*, April 2023. doi: 10.48550/arXiv.2304.00740.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. Technical report, Epoch, March 2024. URL http://arxiv.org/abs/2403.05812.

Daniel E Ho, Jennifer King, Russell C Wald, and Christopher Wan. Building a national AI research resource: A blueprint for the national research cloud. Technical report, Stanford University Human-Centered Artificial Intelligence, 2021.

Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, and Duncan Snidal. International institutions for advanced AI. *arXiv: 2307.04699 [cs.CY]*, July 2023. URL http://arxiv.org/abs/2307.04699.

Mia Hoffmann and Heather Frase. Adding structure to AI harm. Technical report, Center for Security and Emerging Technology Publications (CSET), July 2023. URL https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. *arXiv: 2402.19464 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.19464.

Sara Hooker. On the limitations of compute thresholds as a governance strategy. *arXiv: 2407.05694 [cs.AI]*, July 2024. URL http://arxiv.org/abs/2407.05694.

Shohreh Hosseinzadeh, Bernardo Sequeiros, Pedro R M Inácio, and Ville Leppänen. Recent trends in applying TPM to cloud computing. *Security and privacy*, 3(1), 2019. ISSN 2475-6725. URL https://onlinelibrary.wiley.com/doi/10.1002/spy2.93.

House of Commons Science, Innovation and Technology Committee. Governance of artificial intelligence (AI). Technical report, House of Commons, 2024. URL https://committees.parliament.uk/publications/45145/documents/223578/default/.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models. *arXiv: 2405.16833 [cs.LG]*, May 2024. URL http://arxiv.org/abs/2405.16833.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Hp. Gradient Cuff: Detecting jailbreak attacks in large language models by exploring refusal loss landscapes. *arXiv: 2403.00867 [cs.CR]*, March 2024. URL http://arxiv.org/abs/2403.00867.

Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, Ashley Llorens, John Langford, Hoi Vo, Li Fei-Fei, Katsu Ikeuchi, and Jianfeng Gao. Position paper: Agent AI towards a holistic intelligence. *arXiv: 2403.00833 [cs.AI]*, February 2024a. URL http://arxiv.org/abs/2403.00833.

Raffaele Huang. The underground network sneaking nvidia chips into china, July 2024. URL https://www.wsj.com/tech/the-underground-network-sneaking-nvidia-chips-into-china-f733aaa6.

Tiansheng Huang, Sihao Hu, Faith Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv: 2409.01586 [cs.CL]*, September 2024b. URL http://arxiv.org/abs/2409.01586.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language model. *arXiv: 2402.01109 [cs.LG]*, February 2024c. URL http://arxiv.org/abs/2402.01109.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv: 2401.05566 [cs.CR]*, January 2024. doi: 10.48550/arXiv.2401.05566.

Hugging Face. Gated datasets, 2024. URL https://huggingface.co/docs/hub/en/datasets-gated. Accessed: 2024-7-17.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 1859–1876, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://dl.acm.org/doi/10.1145/3531146.3533233.

Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static AI evaluations: Advancing human interaction evaluations for LLM harms and risks. *arXiv: 2405.10632 [cs.CY]*, May 2024. URL http://arxiv.org/abs/2405.10632.

IEEE. IEEE draft guide: Adoption of the project management institute (PMI) standard: A guide to the project management body of knowledge (PMBOK guide)-2008 (4th edition), June 2011. URL http://dx.doi.org/10.1109/IEEESTD.2011.5937011.

IETF Datatracker. Remote ATtestation ProcedureS (rats), 2024. URL https://datatracker.ietf.org/wg/rats/about/. Accessed: 2024-7-17.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv: 2202.00622 [stat.ML]*, February 2022. URL http://arxiv.org/abs/2202.00622.

Vincent Immler, Johannes Obermaier, Martin König, Matthias Hiller, and Georg Sig. B-TREPID: Batteryless tamper-resistant envelope with a PUF and integrity detection. In *2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pp. 49–56. IEEE, April 2018. ISBN 9781538647318, 9781538647325. URL http://dx.doi.org/10.1109/HST.2018.8383890.

Vincent Immler, Johannes Obermaier, Kuan Kuan Ng, Fei Xiang Ke, Jinyu Lee, Yak Peng Lim, Wei Koon Oh, Keng Hoong Wee, and Georg Sigl. Secure physical enclosures from covers with Tamper-Resistance. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 51–96, 2019. ISSN 2569-2925, 2569-2925. URL https://tches.iacr.org/index.php/TCHES/article/view/7334.

Intel. Intel converged security and management engine (intel CSME) security. Technical Report 631900, Intel, 2022.

International Organization for Standardization. ISO/IEC 22237-1:2021: Information technology — data centre facilities and infrastructures — part 1: General concepts. Technical report, ISO/IEC, 2021. URL https://www.iso.org/standard/78550.html.

International Organization for Standardization. ISO/IEC 23894:2023: Information technology — artificial intelligence — guidance on risk management. Technical report, ISO/IEC, 2023. URL https://www.iso.org/standard/77304.html.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In C Maria Keet, Hung-Yi Lee, and Sina Zarrieß (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.inlg-main.3.

Ahmed Irfan, Kyle D Julian, Haoze Wu, Clark Barrett, Mykel J Kochenderfer, Baoluo Meng, and James Lopez. Towards verification of neural networks for small unmanned aircraft collision avoidance. In *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*. IEEE, 2020. URL http://dx.doi.org/10.1109/DASC50938.2020.9256616.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20)*, pp. 1345–1362, USA, August 2020. USENIX Association. ISBN 9781939133175.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-Yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv: 2309.00614 [cs.LG]*, September 2023a. doi: 10.48550/arXiv.2309.00614.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv: 2311.12786 [cs.LG]*, November 2023b. URL http://arxiv.org/abs/2311.12786.

Shrey Jain, Zoë Hitzig, and Pamela Mishkin. Contextual confidence and generative AI. *arXiv: 2311.01193 [cs.AI]*, November 2023c. URL http://arxiv.org/abs/2311.01193.

Markus Jakobsson and Ari Juels. Proofs of work and bread pudding Protocols(Extended abstract). In Bart Preneel (ed.), *Secure Information Networks: Communications and Multimedia Security IFIP TC6/TC11 Joint Working Conference on Communications and Multimedia Security (CMS'99) September 20–21, 1999, Leuven, Belgium*, pp. 258–272. Springer US, Boston, MA, 1999. ISBN 9780387355689. URL https://doi.org/10.1007/978-0-387-35568-9_18.

Olli Järviniemi and Evan Hubinger. Uncovering deceptive tendencies in language models: A simulated company AI assistant. *arXiv: 2405.01576 [cs.CL]*, April 2024. URL http://arxiv.org/abs/2405.01576.

Julian Jaursch, Jakob Ohme, and Ulrike Klinger. Enabling research with publicly accessible platform data: Early DSA compliance issues and suggestions for improvement. Technical report, Weizenbaum Institut, 2024. URL https://www.weizenbaum-library.de/handle/id/572.

Maha Jebalia, Asma Ben Letaïfa, Mohamed Hamdi, and Sami Tabbane. A fair resource allocation approach in cloud computing environments. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 54–57. IEEE, June 2018. ISBN 9781538669167, 9781538669174. URL http://dx.doi.org/10.1109/WETICE.2018.00017.

Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-Learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056. IEEE, May 2021. ISBN 9781728189345, 9781728189352. URL http://dx.doi.org/10.1109/SP40001.2021.00106.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos. Towards efficient data valuation based on the shapley value. *arXiv: 1902.10275 [cs.LG]*, February 2019. URL http://arxiv.org/abs/1902.10275.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv: 2401.06059 [cs.CL]*, January 2024. URL http://arxiv.org/abs/2401.06059.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in english: Cross-Lingual evaluation of large language models for healthcare queries. *arXiv: 2310.13132 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.13132.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15307–15329. PMLR, 2023. URL https://proceedings.mlr.press/v202/jones23a.html.

kaggle. Code competition FAQ, 2024. URL https://www.kaggle.com/docs/competitions. Accessed: 2024-7-17.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G L D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv: 1912.04977 [cs.LG]*, December 2019. URL http://arxiv.org/abs/1912.04977.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10697–10707. PMLR, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models. *arXiv: 2403.07918 [cs.CY]*, February 2024a. doi: 10.48550/arXiv.2403.07918.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *arXiv: 2407.01502 [cs.LG]*, July 2024b. URL http://arxiv.org/abs/2407.01502.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification*, pp. 97–117. Springer International Publishing, 2017a. URL http://dx.doi.org/10.1007/978-3-319-63387-9_5.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Towards proving the adversarial robustness of deep neural networks. *arXiv: 1709.02802 [cs.LG]*, September 2017b. URL http://arxiv.org/abs/1709.02802.

Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L Dill, Mykel J Kochenderfer, and Clark Barrett. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification*, pp. 443–452. Springer International Publishing, 2019. URL http://dx.doi.org/10.1007/978-3-030-25540-4_26.

Bryan Kelly, Andrés Lagar-Cavilla, Jeff Andersen, Prabhu Jayana, Piotr Kwidzinski, Rob Strong, John Traver, Louis Ferraro, Ishwar Agarwal, Anjana Parthasarathy, Bharat Pillilli, Vishal Soni, Marius Schilder, Sudhir Mathane, Nathan Nadarajah, and Kor Nielsen. Caliptra: A datacenter system on a chip (SOC) root of trust (RoT). Technical report, Open Compute Project, 2022.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv: 2310.08491 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.08491.

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating Language-Model agents on realistic autonomous tasks. *arXiv: 2312.11671 [cs.CL]*, December 2023. URL http://arxiv.org/abs/2312.11671.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17061–17084. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Jon Kleinberg. Inherent Trade-Offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '18)*, pp. 40, New York, NY, USA, June 2018. Association for Computing Machinery. ISBN 9781450358460. URL https://doi.org/10.1145/3219617.3219634.

Alistair Knott, Dino Pedreschi, Raja Chatila, Tapabrata Chakraborti, Susan Leavy, Ricardo Baeza-Yates, David Eyers, Andrew Trotman, Paul D Teal, Przemyslaw Biecek, Stuart Russell, and Yoshua Bengio. Generative AI models should include detection mechanisms as a condition for public release. *Ethics and information technology*, 25(4):55, October 2023. ISSN 1388-1957, 1572-8439. URL https://doi.org/10.1007/s10676-023-09728-4.

Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, and Mark Ibrahim. CrypTen: Secure Multi-Party Computation Meets Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4961–4973. Curran Associates, Inc., 2021. URL https://papers.neurips.cc/paper/2021/hash/2754518221cfbc8d25c13a06a4cb8421-Abstract.html.

Noam Kolt. Algorithmic black swans. *Washington University Law Review*, 101, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4370566.

Noam Kolt. Governing AI agents, 2024. URL https://www.ssrn.com/abstract=4772956.

Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B Sandbrink, and Thomas Woodside. Responsible reporting for frontier AI development. *arXiv: 2404.02675 [cs.CY]*, April 2024. doi: 10.48550/arXiv.2404.02675.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni,

Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F P Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of Web-Crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. ISSN 2307-387X. URL https://aclanthology.org/2022.tacl-1.4.

Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J D Vermeer, and Zev Winkelman. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. RAND Corporation, Santa Monica, CA, 2024. URL http://dx.doi.org/10.7249/WRA3056-1.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying LLM safety against adversarial prompting. *arXiv: 2309.02705 [cs.CL]*, September 2023. doi: 10.48550/arXiv.2309.02705.

Nishant Kumar, Mayank Rathee, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. CrypTFlow: Secure TensorFlow inference. *arXiv: 1909.07814 [cs.CR]*, September 2019. URL http://arxiv.org/abs/1909.07814.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-Based constrained sampling from language models. *arXiv: 2205.12558 [cs.CL]:*, May 2022. URL http://arxiv.org/abs/2205.12558.

Tsai-Chi Kuo, Chien-Yun Kuo, and Liang-Wei Chen. Assessing environmental impacts of nanoscale semiconductor manufacturing from the life cycle assessment perspective. *Resources, Conservation and Recycling*, 182:106289, July 2022. ISSN 0921-3449. URL https://www.sciencedirect.com/science/article/pii/S0921344922001379.

Lindsey Kuper, Guy Katz, Justin Gottschlich, Kyle Julian, Clark Barrett, and Mykel Kochenderfer. Toward scalable verification for Safety-Critical deep networks. *arXiv: 1801.05950 [cs.AI]*, January 2018. URL http://arxiv.org/abs/1801.05950.

Sabrina Küspert, Nicolas Moës, and Connor Dunlop. The value chain of general-purpose AI, 2023. URL https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/. Accessed: 2024-7-17.

Seth Lazar. Frontier AI ethics: Anticipating and evaluating the societal impacts of generative agents. *arXiv: 2404.06750 [cs.CY]*, April 2024. URL http://arxiv.org/abs/2404.06750.

Seth Lazar and Alondra Nelson. Ai safety on whose terms?, 2023.

Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin' 'bout AI generation: Copyright and the Generative-AI supply chain (the short version). In *Proceedings of the Symposium on Computer Science and Law (CSLAW '24)*, pp. 48–63, New York, NY, USA, March 2024. Association for Computing Machinery. ISBN 9798400703331. URL https://doi.org/10.1145/3614407.3643696.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S Bernstein, and Percy Liang. Evaluating Human-Language model interaction, July 2023a. URL https://openreview.net/pdf?id=hjDYJUn9l1.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of Text-To-Image models. *arXiv: 2311.04287 [cs.CV]*, November 2023b. URL http://arxiv.org/abs/2311.04287.

Paddy Leerssen. Platform research access in article 31 of the digital services act: Sword without a shield? *Verfassungsblog: On Matters Constitutional*, 2021. ISSN 2366-7044. URL https://intr2dok.vifa-recht.de/receive/mir_mods_00011130.

Paddy Leerssen. *Seeing what others are seeing: Studies in the regulation of transparency for social media recommender systems*. PhD thesis, Faculty of Law, 2023.

Paddy Leerssen, Amélie P Heldt, and Matthias C Kettemann. Scraping By? Europe's law and policy on social media research access. In Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe (eds.), *Challenges and Perspectives of Hate Speech Research*, volume 12, pp. 405–425. Digital Communication Research, Berlin, 2023. ISBN 9783945681121. URL http://dx.doi.org/10.48541/dcr.v12.24.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA fine-tuning efficiently undoes safety training in llama 2-chat 70B. *arXiv: 2310.20624 [cs.LG]*, October 2023. URL http://arxiv.org/abs/2310.20624.

David Leslie, Cami Rincon, Morgan Briggs, Antonella Perini, Smera Jayadeva, Ann Borda, S J Bennett, Christopher Burr, Mhairi Aitken, Michael Katell, Claudia Fischer, Janis Wong, and Ismael Kherroubi Garcia. AI fairness in practice. *arXiv: 2403.14636 [cs.CY]*, February 2024. URL http://arxiv.org/abs/2403.14636.

Haodong Li, Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, Yang Liu, Guoai Xu, Guosheng Xu, and Haoyu Wang. Digger: Detecting copyright content mis-usage in large language model training. *arXiv: 2401.00676 [cs.CR]*, January 2024a. URL http://arxiv.org/abs/2401.00676.

Maximilian Li, Xander Davies, and Max Nadeau. Circuit breaking: Removing model behaviors with targeted ablation. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, 2023a. PMLR. URL http://arxiv.org/abs/2309.05973.

Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Conference on Security Symposium*, Broomfield, CO, 2014. USENIX Association. ISBN 9781931971164. URL https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-li_mu.pdf.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A Hunt, Justin Tienken-Harder, Kevin Y Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv: 2403.03218 [cs.LG]*, March 2024b. doi: 10.48550/arXiv.2403.03218.

Xiaoguo Li, Bowen Zhao, Guomin Yang, Tao Xiang, Jian Weng, and Robert H Deng. A survey of secure computation using trusted execution environments. *arXiv: 2302.12150 [cs.CR]*, February 2023b. URL http://arxiv.org/abs/2302.12150.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang,

Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, February 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW.

Jinkun Lin, Anqi Zhang, Mathias Lecuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the effect of training data on deep learning predictions via randomized experiments. *arXiv: 2206.10013 [cs.LG]*, June 2022a. URL http://arxiv.org/abs/2206.10013.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.229.

Simon Lindgren and Virginia Dignum. Beyond AI solutionism: toward a multi-disciplinary approach to artificial intelligence in society. In *Handbook of Critical Studies of Artificial Intelligence*, pp. 163–172. Edward Elgar Publishing, November 2023. ISBN 9781803928562. URL https://www.elgaronline.com/edcollchap/book/9781803928562/book-part-9781803928562-19.xml?tab_body=abstract-copy1.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip S Yu. A survey of text watermarking in the era of large language models. *arXiv: 2312.07913 [cs.CL]*, December 2023a. URL http://arxiv.org/abs/2312.07913.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *arXiv: 2402.08787 [cs.LG]*, February 2024a. doi: 10.48550/arXiv.2402.08787.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *arXiv: 2310.04451 [cs.CL]*, October 2023b. URL http://arxiv.org/abs/2310.04451.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. *arXiv: 2402.10058 [cs.CL]*, February 2024b. doi: 10.48550/arXiv.2402.10058.

Orly Lobel. The ai regulatory pyramid: A taxonomy & analysis of the emerging toolbox in the global race for the regulation and governance of artificial intelligence. *Loy. LAL Rev.*, 57:859, 2024.

Andrew Lohn. Hacking AI: A primer for policymakers on machine learning cybersecurity. Technical report, {Center for Security and Emerging Technology}, December 2020. URL https://cset.georgetown.edu/publication/hacking-ai/.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv: 2310.16787 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.16787.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. A safe harbor for AI evaluation and red teaming. *arXiv: 2403.04893 [cs.AI]*, March 2024a. doi: 10.48550/arXiv.2403.04893.

Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara, and Sandy Pentland. Data authenticity, consent, and provenance for AI are all broken: What will it take to fix them? *An MIT Exploration of Generative AI*, March 2024b. URL https://mit-genai.pubpub.org/pub/uk7op8zs.

Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 182–189, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.24/.

Alexandra Sasha Luccioni and Alex Hernandez-Garcia. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv: 2302.08476 [cs.LG]*, February 2023. URL http://arxiv.org/abs/2302.08476.

Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 199–212, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://doi.org/10.1145/3531146.3533086.

Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of AI deployment? *arXiv: 2311.16863 [cs.LG]*, November 2023a. URL http://arxiv.org/abs/2311.16863.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of machine learning research*, 24(253):1–15, 2023b. ISSN 1532-4435, 1533-7928. URL https://jmlr.org/papers/v24/23-0069.html.

Sasha Luccioni. Energy star ratings for AI models, 2024. URL https://huggingface.co/blog/sasha/energy-star-ai-proposal. Accessed: 2024-7-18.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv: 2402.16835 [cs.CL]*, February 2024. doi: 10.48550/arXiv.2402.16835.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv: 2406.10229 [cs.LG]*, June 2024. URL http://arxiv.org/abs/2406.10229.

Gary Marcus and Reid Southen. Generative AI has a visual plagiarism problem, January 2024. URL https://spectrum.ieee.org/midjourney-copyright. Accessed: 2024-7-14.

Helen Margetts. Rethinking AI for good governance. *Daedalus*, 151(2):360–371, May 2022. ISSN 0011-5266, 1548-6192. URL https://direct.mit.edu/daed/article-pdf/151/2/360/2060573/daed_a_01922.pdf.

Helen Margetts and Cosmina Dorobantu. Rethink government with AI. *Nature*, pp. 163–165, April 2019. URL http://dx.doi.org/10.1038/d41586-019-01099-5.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A Meta-Evaluation of 769 papers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7297–7306, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.566.

Miljan Martic, Jan Leike, Andrew Trask, Matteo Hessel, Shane Legg, and Pushmeet Kohli. Scaling shared model governance via model splitting. *arXiv: 1812.05979 [cs.LG]*, December 2018. URL http://arxiv.org/abs/1812.05979.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Combining generative artificial intelligence (AI) and the internet: Heading towards evolution or degradation? *arXiv: 2303.01255 [cs.CV]*, February 2023. URL http://arxiv.org/abs/2303.01255.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The AI index 2024 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, USA, April 2024. URL https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv: 2402.04249 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.04249.

Louise McCormack and Malika Bendechache. Ethical ai governance: Methods for evaluating trustworthy ai. *arXiv preprint arXiv:2409.07473*, 2024.

Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *arXiv: 2011.08512 [cs.CY]*, November 2020. URL http://arxiv.org/abs/2011.08512.

Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv: 2402.09880 [cs.AI]*, February 2024. doi: 10.48550/arXiv.2402.09880.

Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. FLIRT: Feedback loop in-context red teaming. *arXiv: 2308.04265 [cs.AI]*, August 2023. URL http://arxiv.org/abs/2308.04265.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, Virtual, October 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 735–746, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 9781450383097. URL https://dl.acm.org/doi/10.1145/3442188.3445935.

Katina Michael, Roba Abbas, Rafael A Calvo, George Roussos, Eusebio Scornavacca, and Samuel Fosso Wamba. Manufacturing consent: The modern pandemic of technosolutionism. *IEEE Transactions on Technology and Society*, 1(2):68–72, June 2020. ISSN 2637-6415. URL http://dx.doi.org/10.1109/TTS.2020.2994381.

Microsoft. Governing AI: A blueprint for the future. Technical report, Microsoft, 2023. URL https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.

Fraser Mince, Dzung Dinh, Jonas Kgomo, Neil Thompson, and Sara Hooker. The grand illusion: The myth of software portability and implications for ML progress. *arXiv: 2309.07181 [cs.SE]*, September 2023. URL http://arxiv.org/abs/2309.07181.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv: 2110.11309 [cs.LG]*, October 2021. URL http://arxiv.org/abs/2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-Based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 15817–15831. PMLR, June 2022a. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pp. 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 9781450361255. URL https://doi.org/10.1145/3287560.3287596.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv: 2212.05129 [cs.AI]*, December 2022b. doi: 10.48550/arXiv.2212.05129.

Fan Mo, Zahra Tarkhani, and Hamed Haddadi. Machine learning with confidential computing: A systematization of knowledge. *ACM computing surveys*, 56(11):1–40, November 2024. ISSN 0360-0300, 1557-7341. URL https://dl.acm.org/doi/10.1145/3670007.

Nicolas Moës and Frank Ryan. Heavy is the head that wears the crown: A risk based tiered approach to governing general purpose AI. Technical report, The Future Society„ 2023.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, May 2023. ISSN 2730-5961. URL https://doi.org/10.1007/s43681-023-00289-2.

Christopher Morten, Gabriel Nicholas, and Salome Viljoen. Researcher access to social media data: Lessons from clinical trial data sharing. *Berkeley Technology Law Journal*, 39(109), 2024. ISSN 1556-5068. URL https://lawcat.berkeley.edu/record/1288021.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. Assembling accountability: Algorithmic impact assessment for the public interest. Technical report, Data & Society, June 2021. URL https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. URL https://aclanthology.org/2022.tacl-1.6.

Christopher A Mouton, Caleb Lucas, and Ella Guest. The operational risks of AI in Large-Scale biological attacks: A Red-Team approach. Technical report, RAND Corporation, October 2023. URL https://www.rand.org/pubs/research_reports/RRA2977-1.html.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation, October 2023. URL https://openreview.net/pdf?id=AKJLnDgzkm.

MULTIBEAM. Applications, 2024. URL https://multibeamcorp.com/applications/. Accessed: 2024-7-17.

Antonio Muñoz, Ruben Ríos, Rodrigo Román, and Javier López. A survey on the (in)security of trusted execution environments. *Computers & Security*, 129:103180, June 2023. ISSN 0167-4048. URL https://www.sciencedirect.com/science/article/pii/S0167404823000901.

Micah Musser, Rebecca Gelles, Catherine Aiken, and Andrew Lohn. "the main resource is the human": A survey of AI researchers on the importance of compute. Technical report, Center for Security and Emerging Technology, April 2023. URL https://cset.georgetown.edu/publication/the-main-resource-is-the-human/.

David Mytton. Data centre water consumption. *npj Clean Water*, 4(1):1–6, February 2021. ISSN 2059-7037, 2059-7037. URL https://www.nature.com/articles/s41545-021-00101-w.

Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 40–48. IEEE, June 2019. ISBN 9781728125060, 9781728125077. URL http://dx.doi.org/10.1109/CVPRW.2019.00011.

Arvind Narayanan and Sayash Kapoor. AI safety is not a model property, 2024. URL https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

Arvind Narayanan, Sayash Kapoor, and Seth Lazar. Model alignment protects against accidental harms, not intentional ones, 2023. URL https://www.aisnakeoil.com/p/model-alignment-protects-against. Accessed: 2024-7-15.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv: 2311.17035 [cs.LG]*, November 2023. doi: 10.48550/arXiv.2311.17035.

National Artificial Intelligence Research Resource Task Force. Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource. Technical report, NAIRR, 2023.

National Institute of Standards and Technology (NIST). Biden-Harris administration announces new NIST public working group on AI, June 2023. URL https://www.nist.gov/news-events/news/2023/06/biden-harris-administration-announces-new-nist-public-working-group-ai. Accessed: 2024-7-14.

National Institute of Standards and Technology (NIST). Biden-Harris administration announces First-Ever consortium dedicated to AI safety, February 2024. URL https://www.nist.gov/news-events/news/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated-ai. Accessed: 2024-7-16.

NDIF proposal. National deep inference facility for very large language models (NDIF): Project proposal to the NSF, 2023. URL https://thevisible.net/docs/NDIF-proposal.pdf.

Sree Harsha Nelaturu, Nishaanth Kanna Ravichandran, Cuong Tran, Sara Hooker, and Ferdinando Fioretto. On the fairness impacts of hardware selection in machine learning. *arXiv: 2312.03886 [cs.LG]*, December 2023. URL http://arxiv.org/abs/2312.03886.

Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RAND Corporation, Santa Monica, CA, 2024. URL http://dx.doi.org/10.7249/RRA2849-1.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv: 2209.02299 [cs.LG]*, September 2022. URL http://arxiv.org/abs/2209.02299.

NHS Research SDE Network. Secure data environment, 2024. URL https://digital.nhs.uk/services/secure-data-environment-service. Accessed: 2024-7-17.

Gabriel Nicholas. Grounding ai policy: Towards researcher access to ai usage data. Technical report, Center for Democracy & Technology, August 2024. URL https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.

Robert C Nickerson, Upkar Varshney, and Jan Muntermann. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359, 2013.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16805–16827. PMLR, June 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

NIST. NIST AIRC - Crosswalk Documents, 2023a. URL https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks. Accessed: 2024-7-18.

NIST. NIST AI RMF playbook, 2023b. URL https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook. Accessed: 2024-7-18.

NIST. Artificial intelligence risk management framework: Generative artificial intelligence profile, 2024. URL https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

Johannes Obermaier and Vincent Immler. The past, present, and future of physical security enclosures: From Battery-Backed monitoring to PUF-Based inherent security and beyond. *Journal of Hardware and Systems Security*, 2(4):289–296, December 2018. ISSN 2509-3436. URL https://doi.org/10.1007/s41635-018-0045-2.

Joe O'Brien, Shaun Ee, and Zoe Williams. Deployment corrections: An incident response framework for frontier AI models. Technical report, Institute for AI Policy and Strategy, 2023. URL https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651c397fc04af033499df9f8/1696348544356/Deployment+corrections_+an+incident+response+framework+for+frontier+AI+models.pdf.

Serena Oduro and Tamara Kneese. AI governance needs sociotechnical expertise: Why the humanities and social sciences are critical to government efforts. Technical report, Data&society, 2024. URL https://datasociety.net/wp-content/uploads/2024/05/DS_AI_Governance_Policy_Brief.pdf.

OECD. Oecd principles on artificial intelligence. *Organisation for Economic Co-Operation and Development*, 2019. URL https://oecd.ai/en/ai-principles.

OECD. Stocktaking for the development of an AI incident definition. Technical report, OECD, October 2023. URL https://www.oecd-ilibrary.org/science-and-technology/stocktaking-for-the-development-of-an-ai-incident-definition_c323ac71-en.

OECD.AI Policy Observatory. OECD AI incidents monitor, 2024. URL https://oecd.ai/en/incidents-methodology. Accessed: 2024-7-18.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *arXiv: 2402.17861 [cs.CY]*, February 2024. doi: 10.48550/arXiv.2402.17861.

Chris Olah. Interpretability dreams, 2023. URL https://transformer-circuits.pub/2023/interpretability-dreams/index.html. Accessed: 2024-7-16.

OpenAI. Preparedness, 2024a. URL https://openai.com/preparedness/. Accessed: 2024-5-26.

OpenAI. Reimagining secure infrastructure for advanced AI, 2024b. URL https://openai.com/index/reimagining-secure-infrastructure-for-advanced-ai/. Accessed: 2024-7-17.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo,

Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. Technical report, OpenAI, March 2024. URL http://arxiv.org/abs/2303.08774.

Tribhuvanesh Orekondy, B Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of Black-Box models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4949–4958, December 2018. ISSN 1063-6919, 2575-7075. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Orekondy_Knockoff_Nets_Stealing_Functionality_of_Black-Box_Models_CVPR_2019_paper.pdf.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *arXiv: 2310.17623 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.17623.

Organisation for Economic Co-Operation and Development. A blueprint for building national compute capacity for artificial intelligence. Technical report, OECD, February 2023. URL https://www.oecd-ilibrary.org/science-and-technology/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: A task-oriented investigation of User-GPT interactions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2375–2393, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.146.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22, New York, NY, USA, October 2023a. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), May 2024. ISSN 2666-3899. URL https://cell.com/patterns/retrieve/pii/S266638992400103X?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS266638992400103X%3Fshowall%3Dtrue.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. *arXiv: 2303.14186 [stat.ML]*, March 2023b. URL http://arxiv.org/abs/2303.14186.

Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Comput. Surv.*, December 2023. ISSN 0360-0300. URL https://doi.org/10.1145/3637549.

Partnership on AI. PAI's guidance for safe foundation model deployment, October 2023. URL https://partnershiponai.org/modeldeployment/. Accessed: 2024-7-18.

Dylan Patel. China AI & semiconductors rise: US sanctions have failed, 2023. URL https://www.semianalysis.com/p/china-ai-and-semiconductors-rise. Accessed: 2024-7-17.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, July 2022. ISSN 0018-9162, 1558-0814. URL https://ieeexplore.ieee.org/document/9810097.

David Patterson, Jeffrey M Gilbert, Marco Gruteser, Efren Robles, Krishna Sekar, Yong Wei, and Tenghui Zhu. Energy and emissions of machine learning on smartphones vs. the cloud. *Communications of the ACM*, 67(2):86–97, January 2024. ISSN 0001-0782. URL https://doi.org/10.1145/3624719.

Michael Quinn Patton. *Qualitative research & evaluation methods*. sage, 2002.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv: 2406.17557 [cs.CL]*, June 2024. URL http://arxiv.org/abs/2406.17557.

Shengyun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv: 2405.17374 [cs.LG]*, May 2024. URL http://arxiv.org/abs/2405.17374.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.225.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with Model-Written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.847.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana

Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating frontier models for dangerous capabilities. Technical report, Google Deepmind, March 2024. URL http://arxiv.org/abs/2403.13793.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, Shengyun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. LLM self defense: By self examination, LLMs know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024*, Vienna, Austria, March 2024. URL https://openreview.net/forum?id=YoqgcIA19o.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. The ROOTS search tool: Data transparency for LLMs. *arXiv: 2302.14035 [cs.CL]*, February 2023. URL http://arxiv.org/abs/2302.14035.

Hadrien Pouget. The EU's AI act is barreling toward AI standards that do not exist, 2023. URL https://www.lawfaremedia.org/article/eus-ai-act-barreling-toward-ai-standards-do-not-exist. Accessed: 2024-7-18.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. On the challenges of using Black-Box APIs for toxicity evaluation in research. *arXiv: 2304.12397 [cs.CL]*, April 2023. URL http://arxiv.org/abs/2304.12397.

Usvsn Sai Prashanth, Alvin Deng, Kyle O'Brien, S V Jyothir, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. *arXiv: 2406.17746 [cs.CL]*, June 2024. URL http://arxiv.org/abs/2406.17746.

Presidency of the Council of the European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - analysis of the final compromise text with a view to agreement. Technical Report 5662/24 LIMITE, Council of the European Union, 2024. URL https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

Project Oak. Project oak, 2024. URL https://github.com/project-oak/oak.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1776–1826, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://doi.org/10.1145/3531146.3533231.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The 12th International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, October 2023. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv: 2406.05946 [cs.CR]*, June 2024a. URL http://arxiv.org/abs/2406.05946.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight LLMs. *arXiv: 2412.07097 [cs.CR]*, December 2024b. URL http://arxiv.org/abs/2412.07097.

Zhenting Qi, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. Follow my instruction and spill the beans: Scalable data extraction from Retrieval-Augmented generation systems. *arXiv: 2402.17840 [cs.CL]*, February 2024c. URL http://arxiv.org/abs/2402.17840.

Jenny Quang. Does training AI violate copyright law? *Boalt Hall School of Law, University of California, Berkeley*, 36(4):1407, 2021. ISSN 1086-3818. URL https://btlj.org/wp-content/uploads/2023/02/0003-36-4Quang.pdf.

Manish Raghavan. *The Societal Impacts of Algorithmic Decision-Making*, volume 53. Association for Computing Machinery, New York, NY, USA, 1 edition, August 2023.

Noorjahan Rahman and Eduardo Santacana. Beyond fair use: Legal risk evaluation for training LLMs on copyrighted text, 2023. URL https://blog.genlaw.org/CameraReady/57.pdf.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track (Round 2)*, Virtual, August 2021. URL https://openreview.net/forum?id=j6NxpQbREA1.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 959–972, 2022a.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, pp. 557–571, New York, NY, USA, July 2022b. Association for Computing Machinery. ISBN 9781450392471. URL https://dl.acm.org/doi/10.1145/3514094.3534181.

Bogdana Rakova and Roel Dobbe. Algorithms as Social-Ecological-Technological systems: An environmental justice lens on algorithmic audits. *arXiv: 2305.05733 [cs.CY]*, May 2023. URL http://arxiv.org/abs/2305.05733.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6034–6055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.405/.

William Alan Reinsch, Matthew Schleich, and Thibault Denamiel. Insight into the U.S. semiconductor export controls update, 2023. URL https://www.csis.org/analysis/insight-us-semiconductor-export-controls-update. Accessed: 2024-7-15.

Anka Reuel and Trond Arne Undheim. Generative AI needs adaptive governance. *arXiv: 2406.04554 [cs.CY]*, June 2024. URL http://arxiv.org/abs/2406.04554.

Anka Reuel, Lisa Soder, Ben Bucknall, and Trond Arne Undheim. Position paper: Technical research and talent is needed for effective AI governance. *arXiv: 2406.06987 [cs.CY]*, June 2024a. URL http://arxiv.org/abs/2406.06987.

Anka Reuel, Lisa Soder, Benjamin Bucknall, and Trond Arne Undheim. Position: Technical research and talent is needed for effective AI governance. In *Forty-first International Conference on Machine Learning*, June 2024b. URL https://openreview.net/pdf?id=Be2B6f0ps1.

M Sadegh Riazi, Christian Weinert, Oleksandr Tkachenko, Ebrahim M Songhori, Thomas Schneider, and Farinaz Koushanfar. Chameleon: A hybrid secure computation framework for machine learning applications. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, pp. 707–721, New York, NY, USA, May 2018. Association for Computing Machinery. ISBN 9781450355766. URL https://doi.org/10.1145/3196494.3196522.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/pdf?id=m2NVG4Htxs.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising effectively prevents harmful fine-tuning on LLMs. *arXiv: 2405.14577 [cs.CL]*, May 2024a. URL http://arxiv.org/abs/2405.14577.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv: 2402.16382 [cs.CL]*, February 2024b. URL http://arxiv.org/abs/2402.16382.

Marcello Ruberti. The chip manufacturing industry: Environmental impacts and eco-efficiency analysis. *The Science of the total environment*, 858(Pt 2):159873, February 2023. ISSN 0048-9697, 1879-1026. URL http://dx.doi.org/10.1016/j.scitotenv.2022.159873.

Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv: 1811.04017 [cs.LG]*, November 2018. URL http://arxiv.org/abs/1811.04017.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of AI-Image detectors: Fundamental limits and practical attacks. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/pdf?id=dLoAdIKENc.

Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, volume 1, pp. 57–64. IEEE, 2015.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated text be reliably detected? *arXiv: 2303.11156 [cs.CL]*, March 2023. doi: 10.48550/arXiv.2303.11156.

Olawale Salaudeen and Moritz Hardt. ImageNot: A contrast with ImageNet preserves model rankings. *arXiv: 2404.02112 [cs.LG]*, April 2024. URL http://arxiv.org/abs/2404.02112.

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing power and the governance of artificial intelligence. *arXiv: 2402.08797 [cs.CY]*, February 2024. URL http://arxiv.org/abs/2402.08797.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. ARB: Advanced reasoning benchmark for large language models. *arXiv: 2307.13692 [cs.CL]*, July 2023. URL http://arxiv.org/abs/2307.13692.

Rylan Schaeffer. Pretraining on the test set is all you need. *arXiv: 2309.08632 [cs.CL]*, September 2023. URL http://arxiv.org/abs/2309.08632.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? *arXiv: 2406.04391 [cs.LG]*, June 2024. URL http://arxiv.org/abs/2406.04391.

Paul Scharre. Future-Proofing frontier AI regulation: Projecting future compute for frontier AI models. Technical report, CNAS, March 2024. URL https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, Vienna, Austria, March 2024. URL https://openreview.net/forum?id=HduMpot9sJ.

Jonas Schuett, Markus Anderljung, Alexis Carlier, Leonie Koessler, and Ben Garfinkel. From principles to rules: A regulatory approach for frontier AI. *arXiv: 2407.07300 [cs.CY]*, July 2024. URL http://arxiv.org/abs/2407.07300.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv: 2210.08402 [cs.CV]*, October 2022. URL http://arxiv.org/abs/2210.08402.

Roei Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. You autocomplete me: Poisoning vulnerabilities in neural code completion. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1559–1575, 2021. URL https://www.usenix.org/system/files/sec21-schuster.pdf.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv: 2310.11324 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.11324.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta. Open-Sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Technical report, Centre for the Governance of AI, September 2023. URL http://arxiv.org/abs/2311.09227.

Andrew D Selbst. An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1), 2021.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN 2022)*, pp. 1–8, Padua, Italy, July 2022. URL https://ieeexplore.ieee.org/document/9891914.

Rusheb Shah, Quentin Feuillade Montixi, Soroush Pour, Arush Tagade, and Javier Rando. Scalable and transferable Black-Box jailbreaks for language models via persona modulation. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Socially Responsible Language Modelling Research Workshop (SoLaR)*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=x3Ltqz1UFg.

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *arXiv: 2305.06360 [cs.LG]*, May 2023. URL http://arxiv.org/abs/2305.06360.

Tommy Shaffer Shane. AI incident reporting: Addressing a gap in the UK's regulation of AI. Technical report, The Centre for Long-Term Resilience, June 2024. URL https://www.longtermresilience.org/post/ai-incident-reporting-addressing-a-gap-in-the-uk-s-regulation-of-ai.

ShareGPT. ShareGPT, 2022. URL https://sharegpt.com/. Accessed: 2024-7-17.

Yonadav Shavit. What does it take to catch a chinchilla? Verifying rules on Large-Scale neural network training via compute monitoring. *arXiv: 2303.11341 [cs.LG]*, March 2023. URL http://arxiv.org/abs/2303.11341.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G Robinson. Practices for governing aagentic AI systems. *Research Paper, OpenAI*, 2023.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv: 2310.10844 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.10844.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *arXiv: 2210.05791 [cs.HC]*, October 2022. URL http://arxiv.org/abs/2210.05791.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741, 2023.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks. Technical report, Google DeepMind, September 2023. URL http://arxiv.org/abs/2305.15324.

Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too? *arXiv: 2212.10539 [cs.CL]*, December 2022. URL http://arxiv.org/abs/2212.10539.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The 12th International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, October 2023. URL https://openreview.net/forum?id=zWqr3MQuNs.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. *arXiv: 2407.06460 [cs.CL]*, July 2024. doi: 10.48550/arXiv.2407.06460.

Taylor Shin, Yasaman Razeghi, Robert L Logan, IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.346.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, San Jose, CA, USA, 2017. IEEE. ISBN 9781509055333. URL http://ieeexplore.ieee.org/document/7958568/.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv: 2305.17493 [cs.LG]*, May 2023. URL http://arxiv.org/abs/2305.17493.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. Knowledge unlearning for LLMs: Tasks, methods, and challenges. *arXiv: 2311.15766 [cs.CL]*, November 2023. URL http://arxiv.org/abs/2311.15766.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. Metadata archaeology: Unearthing data subsets by leveraging training dynamics. *arXiv: 2209.10015 [cs.LG]*, September 2022. URL http://arxiv.org/abs/2209.10015.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hhacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 1305–1320, New York, NY, USA, June 2024. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3630106.3658974.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *The 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*, pp. 1–6, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 9781450394772. URL https://dl.acm.org/doi/10.1145/3551624.3555285.

Gregory T Smith. On construct validity: Issues of method and measurement. *Psychological assessment*, 17 (4):396–408, December 2005. ISSN 1040-3590. URL http://dx.doi.org/10.1037/1040-3590.17.4.396.

Kevin B Smith. Typologies, taxonomies, and the benefits of policy classification. *Policy studies journal*, 30 (3):379–395, 2002.

Shoshanna Sofaer. Qualitative methods: what are they and why use them? *Health services research*, 34(5 Pt 2):1101, 1999.

Irene Solaiman. The gradient of generative AI release: Methods and considerations. *arXiv: 2302.04844 [cs.CY]*, February 2023. URL http://arxiv.org/abs/2302.04844.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé, III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative AI systems in systems and society. *arXiv: 2306.05949 [cs.CY]*, June 2023. URL http://arxiv.org/abs/2306.05949.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. Universal adversarial attacks with natural triggers for text classification. *arXiv: 2005.00174 [cs.CL]*, May 2020. URL http://arxiv.org/abs/2005.00174.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. *arXiv: 2402.10260 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.10260.

Stavros Souravlas and Stefanos Katsavounis. Scheduling fair resource allocation policies for cloud computing through flow control. *Electronics*, 8(11):1348, November 2019. ISSN 2079-9292, 2079-9292. URL https://www.mdpi.com/2079-9292/8/11/1348.

Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex 'sandy' Pentland. Verifiable evaluations of machine learning models using zkSNARKs. *arXiv: 2402.02675 [cs.LG]*, February 2024. URL http://arxiv.org/abs/2402.02675.

Siddarth Srinivasan. Detecting AI fingerprints: A guide to watermarking and beyond. Technical report, Brookings, January 2024. URL https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea

Santilli, Andreas Stuhlmüller, Andrew M Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B Tenenbaum, Joshua S Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi

Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, May 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv: 2402.19450 [cs.AI]*, February 2024. URL https://arxiv.org/abs/2402.19450.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. *arXiv: 1706.03691 [cs.LG]*, June 2017. URL http://arxiv.org/abs/1706.03691.

Andrei Stoian, Jordan Frery, Roman Bredehoft, Luis Montero, Celia Kherfallah, and Benoit Chevallier-Mames. Deep neural networks for encrypted inference with TFHE. *arXiv: 2302.10906 [cs.LG]*, February 2023. URL http://arxiv.org/abs/2302.10906.

Vincent J Straub, Deborah Morgan, Jonathan Bright, and Helen Margetts. Artificial intelligence in government: Concepts, standards, and a unified framework. *Government information quarterly*, 40(4): 101881, October 2023. ISSN 0740-624X. URL https://www.sciencedirect.com/science/article/pii/S0740624X23000813.

Milton E Strauss and Gregory T Smith. Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, 5:1–25, 2009. ISSN 1548-5943, 1548-5951. URL http://dx.doi.org/10.1146/annurev.clinpsy.032408.153639.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1355.

Jingtong Su, Julia Kempe, and Karen Ullrich. Mission Impossible: A statistical perspective on jailbreaking LLMs. *arXiv: 2408.01420 [cs.LG]*, August 2024. URL http://arxiv.org/abs/2408.01420.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. Detecting personal information in training corpora: an analysis. In Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (eds.), *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pp. 208–220, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.trustnlp-1.18.

Arjun Subramonian, Xingdi Yuan, Hal Daumé, III, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. *arXiv: 2305.09022 [cs.CL]*, May 2023. URL http://arxiv.org/abs/2305.09022.

Haochen Sun, Jason Li, and Hongyang Zhang. zkLLM: Zero knowledge proofs for large language models. *arXiv: 2404.16109 [cs.LG]*, April 2024. URL http://arxiv.org/abs/2404.16109.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. *arXiv: 2408.00761 [cs.LG]*, August 2024. URL http://arxiv.org/abs/2408.00761.

Tan, Knott, Tian, and Wu. CryptGPU: Fast Privacy-Preserving machine learning on the GPU. In *2021 IEEE Symposium on Security and Privacy (SP)*, volume 0, pp. 1021–1038, May 2021. URL http://dx.doi.org/10.1109/SP40001.2021.00098.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

The Allen Institute for Artificial Intelligence. WildChat, 2024. URL https://wildchat.allen.ai/. Accessed: 2024-7-17.

The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Technical report, The White House, October 2023a. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

The White House. FACT SHEET: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, October 2023b. URL https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/. Accessed: 2024-4-23.

The White House Office of Science and Technology Policy. Blueprint for an AI bill of rights: Making automated systems work for the american people. Technical report, White House, November 2023. URL https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

David Thiel. Identifying and eliminating CSAM in generative ML training data and models. Technical report, Stanford Digital Repository, 2023. URL https://purl.stanford.edu/kh752sm9123.

Cheng Ting-Fang. ASML says decoupling chip supply chain is practically impossible. *Financial Times*, 2023. URL https://www.ft.com/content/317be8b3-48d9-411e-b763-261a179c9d0d.

Helen Toner, Jessica Ji, John Bansemer, Lucy Lim, Chris Painter, Courtney Corley, Jess Whittlestone, Matt Botvinick, Mikel Rodriguez, and Ram Shankar Siva Kumar. Skating to where the puck is going: Anticipating and managing risks from frontier AI systems. Technical report, Center for Security and Emerging Technology, 2023. URL http://dx.doi.org/10.51593/2023CA004.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and Fine-Tuned chat models. Technical report, Meta AI, July 2023. URL http://arxiv.org/abs/2307.09288.

Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, Simon Staffell, and José Jaime Villalobos. International governance of civilian AI: A jurisdictional certification approach. Technical report, Oxford Martin AI Governance Initiative, August 2023.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *Proceedings of the 25th USENIX Conference on Security Symposium (SEC'16)*, pp. 601–618, USA, August 2016. USENIX Association. ISBN 9781931971324.

Andrew Trask. Building safe AI: A tutorial for encrypted deep learning, 2017. URL https://iamtrask.github.io/2017/03/17/safe-ai/. Accessed: 2024-7-17.

Andrew Trask, Akshay Sukumar, Antti Kalliokoski, Bennett Farkas, Callis Ezenwaka, Carmen Popa, Curtis Mitchell, Dylan Hrebenach, George-Cristian Muraru, Ionesio Junior, Irina Bejan, Ishan Mishra, Ivoline Ngong, Jack Bandy, Jess Stahl, Julian Cardonnet, Kellye Trask, Kellye Trask, Khoa Nguyen, Kien Dang, Koen van der Veen, Kyoko Eng, Lacey Strahm, Laura Ayre, Madhava Jay, Oleksandr Lytvyn, Osam Kyemenu-Sarsah, Peter Chung, Peter Smith, S Rasswanth, Ronnie Falcon, Shubham Gupta, Stephen Gabriel, Teo Milea, Theresa Thoraldson, Thiago Porto, Tudor Cebere, Yash Gorana, and Zarreen Reza. How to audit an AI model owned by someone else (part 1), 2023. URL https://blog.openmined.org/ai-audit-part-1/. Accessed: 2024-7-17.

Alexey Turchin and David Denkenberger. Classification of global catastrophic risks connected with artificial intelligence. *AI & society*, 35(1):147–163, 2018. ISSN 0951-5666, 1435-5655. URL https://doi.org/10.1007/s00146-018-0845-5.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv: 2308.10248 [cs.CL]*, August 2023. doi: 10.48550/arXiv.2308.10248.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H S Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "Zero-Shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv: 2404.04125 [cs.CV]*, April 2024. URL http://arxiv.org/abs/2404.04125.

UK AI Safety Institute. Inspect, 2024. URL https://ukgovernmentbeis.github.io/inspect_ai/. Accessed: 2024-7-15.

UK Research and Innovation. £300 million to launch first phase of new AI research resource, 2023. URL https://www.ukri.org/news/300-million-to-launch-first-phase-of-new-ai-research-resource/. Accessed: 2024-7-17.

U.S. National Telecommunications and Information Administration. Artificial intelligence: Accountability policy report. Technical report, NTIA, 2024. URL https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned Open-Access multilingual language model. *arXiv: 2402.07827 [cs.CL]*, February 2024. URL http://arxiv.org/abs/2402.07827.

Risto Uuk, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl. A taxonomy of systemic risks from general-purpose ai. *arXiv preprint arXiv:2412.07780*, 2024.

Stephan van Schaik, Adam Batori, Alex Seto, Bader AlBassam, Christina Garman, Thomas Yurek, Andrew Miller, Daniel Genkin, Eyal Ronen, and Yuval Yarom. SGX.Fail, 2022. URL https://sgx.fail/. Accessed: 2024-7-17.

Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, U.S. National Institute of Standards and Technology, Gaithersburg, MD, January 2024. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv: 1812.00564 [cs.LG]*, December 2018. URL http://arxiv.org/abs/1812.00564.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? Limits of LLM scaling based on human-generated data. *arXiv: 2211.04325 [cs.LG]*, October 2022. URL http://arxiv.org/abs/2211.04325.

Sameer Wagh, Divya Gupta, and Nishanth Chandran. SecureNN: 3-party secure computation for neural network training. *Proceedings on Privacy Enhancing Technologies*, 2019(3):26–49, July 2019. ISSN 2299-0984. URL https://petsymposium.org/popets/2019/popets-2019-0035.php.

Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. FALCON: Honest-Majority maliciously secure framework for private deep learning. *arXiv: 2004.02229 [cs.CR]:*, April 2020. URL http://arxiv.org/abs/2004.02229.

Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. Trustless audits without revealing data or models. *arXiv: 2404.04500 [cs.CR]*, April 2024. URL http://arxiv.org/abs/2404.04500.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pp. 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1221.

Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP models. *arXiv: 2010.12563 [cs.CL]*, October 2020. URL http://arxiv.org/abs/2010.12563.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*, New Orleans, LA, USA, November 2023a. URL https://openreview.net/forum?id=kaHpo8OZw2.

Wei Wang, Ben Liang, and Baochun Li. Multi-Resource fair allocation in heterogeneous cloud computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 26(10):2822–2835, October 2015. ISSN 1045-9219, 1558-2183. URL http://dx.doi.org/10.1109/TPDS.2014.2362139.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-Tse Huang, Wenxiang Jiao, and Michael R Lyu. All languages matter: On the multilingual safety of large language models. *arXiv: 2310.00905 [cs.CL]*, October 2023b. URL http://arxiv.org/abs/2310.00905.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. *arXiv: 2311.16421 [cs.CL]*, November 2023c. URL http://arxiv.org/abs/2311.16421.

Matthew T Wansley. Regulation of emerging risks. *Vanderbilt Law Review*, 69(2):401, 2016. ISSN 0042-2533. URL https://scholarship.law.vanderbilt.edu/vlr/vol69/iss2/3.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):1–39, December 2023. ISSN 1833-2595, 1833-2595. URL https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00146-z.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=jA235JGM09.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pp. 24824–24837, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. Proving membership in LLM pretraining data via data watermarks. *arXiv: 2402.10892 [cs.CR]*, February 2024. URL http://arxiv.org/abs/2402.10892.

Junyi Wei, Yicheng Zhang, Zhe Zhou, Zhou Li, and Mohammad Abdullah Al Faruque. Leaky DNN: Stealing Deep-Learning model secret with GPU Context-Switching Side-Channel. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 125–137. IEEE, June 2020. ISBN 9781728158099, 9781728158105. URL http://dx.doi.org/10.1109/DSN48063.2020.00031.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://doi.org/10.1145/3531146.3533088.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative AI systems. Technical report, Google Deepmind, October 2023. URL http://arxiv.org/abs/2310.11986.

Laura Weidinger, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, Stevie Bergman, Mikel Rodriguez, Verena Rieser, and William Isaac. STAR: SocioTechnical approach to red teaming language models. *arXiv: 2406.11757 [cs.AI]*, June 2024. URL http://arxiv.org/abs/2406.11757.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. *arXiv: 2109.07445 [cs.CL]*, September 2021. URL http://arxiv.org/abs/2109.07445.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and T Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track*, volume abs/2302.03668, February 2023. URL http://dx.doi.org/10.48550/arXiv.2302.03668.

Drew Westen and Robert Rosenthal. Quantifying construct validity: Two simple measures. *Journal of personality and social psychology*, 84(3):608–618, March 2003. ISSN 0022-3514. URL https://psycnet.apa.org/fulltext/2003-01588-016.pdf.

Wex Definitions Team. joint and several liability, 2023. URL https://www.law.cornell.edu/wex/joint_and_several_liability. Accessed: 2024-7-16.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark, June 2024. URL https://arxiv.org/abs/2406.19314.

Jess Whittlestone and Jack Clark. Why and how governments should monitor AI development. *arXiv: 2108.12427 [cs.CY]*, August 2021. URL http://arxiv.org/abs/2108.12427.

Wikipedia contributors. TIA-942 — Wikipedia, the free encyclopedia, 2023. URL https://en.wikipedia.org/w/index.php?title=TIA-942&oldid=1177253885.

Wu, Li Liu, Yuchen Guo, Guiguang Ding, J Han, Jialie Shen, and Ling Shao. Unsupervised deep video hashing with balanced rotation. *International Joint Conference on Artificial Intelligence*, pp. 3076–3082, August 2017. URL https://www.ijcai.org/proceedings/2017/0429.pdf.

Tim Wu. In regulating A.I., we may be doing too much. and too little. *The New York Times*, November 2023. ISSN 0362-4331, 1553-8095. URL https://www.nytimes.com/2023/11/07/opinion/biden-ai-regulation.html.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 2875–2886, Gateway, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.174.

Pengtao Xie, Misha Bilenko, Tom Finley, Ran Gilad-Bachrach, Kristin Lauter, and Michael Naehrig. CryptoNets: Neural networks over encrypted data. *arXiv: 1412.6181 [cs.LG]*, December 2014. URL http://arxiv.org/abs/1412.6181.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. *arXiv: 2401.12255 [cs.CR]*, January 2024. URL http://arxiv.org/abs/2401.12255.

Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting AI trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 103–120. IEEE, May 2021. ISBN 9781728189345, 9781728189352. URL http://dx.doi.org/10.1109/SP40001.2021.00034.

Xin Xu and Huiqun Yu. A game theory approach to fair and efficient resource allocation in cloud computing. *Mathematical Problems in Engineering*, 2014:1–14, 2014. ISSN 1024-123X, 1563-5147. URL http://www.hindawi.com/journals/mpe/2014/915878/.

Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof : Confidential and certifiable fairness for neural networks. *arXiv: 2402.12572 [cs.LG]*, February 2024. URL https://arxiv.org/abs/2402.12572.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting Safely-Aligned language models. *arXiv: 2310.02949 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.02949.

Andrew C Yao. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pp. 160–164. IEEE, November 1982. URL http://dx.doi.org/10.1109/SFCS.1982.38.

Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pp. 162–167. IEEE, October 1986. ISBN 9780818607400. URL http://dx.doi.org/10.1109/SFCS.1986.25.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv: 2310.10683 [cs.CL]*, October 2023. URL http://arxiv.org/abs/2310.10683.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. *arXiv: 2307.10928 [cs.CL]*, July 2023. URL http://arxiv.org/abs/2307.10928.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv: 2306.13549 [cs.CV]*, June 2023. URL http://arxiv.org/abs/2306.13549.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. Low-Resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=pn83r8V2sv.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: a flexible and expandable family of evaluations for ai models. *arXiv: 2310.17567 [cs.CL]*, October 2023. URL https://arxiv.org/abs/2310.17567.

Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. AI risk categorization decoded (AIR 2024): From government regulations to corporate policies. *arXiv: 2406.17864 [cs.CY]*, June 2024. URL http://arxiv.org/abs/2406.17864.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF protections in GPT-4 via Fine-Tuning. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico, 2024. URL http://arxiv.org/abs/2311.05553.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv: 2311.04378 [cs.LG]*, November 2023. doi: 10.48550/arXiv.2311.04378.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. *arXiv: 2405.00332 [cs.CL]*, May 2024. URL http://arxiv.org/abs/2405.00332.

Rui Zhang, Jian Liu, Yuan Ding, Zhibo Wang, Qingbiao Wu, and Kui Ren. "adversarial examples" for Proof-of-Learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1408–1422. IEEE, May 2022. ISBN 9781665413169, 9781665413176. URL http://dx.doi.org/10.1109/SP46214.2022.9833596.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*, New Orleans, LA, USA, November 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Andy Zhou, Bo Li, and Haohan Wang. Robust prompt optimization for defending language models against jailbreaking attacks. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, Vienna, Austria, April 2024. URL https://openreview.net/forum?id=cSPXIO7min.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your LLM an evaluation benchmark cheater. *arXiv: 2311.01964 [cs.CL]*, November 2023a. URL http://arxiv.org/abs/2311.01964.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Making harmful behaviors unlearnable for large language models. *arXiv: 2311.02105 [cs.LG]*, November 2023b. URL http://arxiv.org/abs/2311.02105.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. PromptBench: Towards evaluating the robustness of Large Language Models on adversarial prompts. *arXiv: 2306.04528 [cs.CL]*, June 2023a. URL https://github.com/microsoft/promptbench.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. AutoDAN: Interpretable Gradient-Based adversarial attacks on large language models. *arXiv: 2310.15140 [cs.CR]*, October 2023b. URL http://arxiv.org/abs/2310.15140.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A Top-Down approach to AI transparency. *arXiv: 2310.01405 [cs.LG]*, October 2023a. doi: 10.48550/arXiv.2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv: 2307.15043 [cs.CL]*, July 2023b. doi: 10.48550/arXiv.2307.15043.

国家互联网信息办公室国家发展和改革委员会教育部等. Interim measures for the management of generative artificial intelligence services, 2023. URL https://www.chinalawtranslate.com/en/generative-ai-interim/.