

Black-Box Access is Insufficient for Rigorous AI Audits

Stephen Casper*
scasper@mit.edu
MIT CSAIL

Carson Ezell*
cezell@college.harvard.edu
Harvard University

Charlotte Siegmann
MIT

Noam Kolt
University of Toronto

Taylor Lynn Curtis
MIT CSAIL

Benjamin Bucknall
Centre for the Governance of AI

Andreas Haupt
MIT

Kevin Wei
Harvard Law School

Jérémy Scheurer
Apollo Research

Marius Hobbhahn
Apollo Research

Lee Sharkey
Apollo Research

Satyapriya Krishna
Harvard University

Marvin Von Hagen
MIT

Silas Alberti
Stanford University

Alan Chan
Mila, Centre for the Governance of AI

Qinyi Sun
MIT

Michael Gerovitch
MIT

David Bau
Northeastern University

Max Tegmark
MIT

David Krueger
University of Cambridge

Dylan Hadfield-Menell
MIT CSAIL

ABSTRACT

External audits of AI systems are increasingly recognized as a key mechanism for AI governance. The effectiveness of an audit, however, depends on the degree of access granted to auditors. Recent audits of state-of-the-art AI systems have primarily relied on *black-box* access, in which auditors can only query the system and observe its outputs. However, *white-box* access to the system’s inner workings (e.g., weights, activations, gradients) allows an auditor to perform stronger attacks, more thoroughly interpret models, and conduct fine-tuning. Meanwhile, *outside-the-box* access to training and deployment information (e.g., methodology, code, documentation, data, deployment details, findings from internal evaluations) allows auditors to scrutinize the development process and design more targeted evaluations. In this paper, we examine the limitations of black-box audits and the advantages of white- and outside-the-box audits. We also discuss technical, physical, and legal safeguards

for performing these audits with minimal security risks. Given that different forms of access can lead to very different levels of evaluation, we conclude that (1) transparency regarding the access and methods used by auditors is necessary to properly interpret audit results, and (2) white- and outside-the-box access allow for substantially more scrutiny than black-box access alone.

CCS CONCEPTS

• **Security and privacy** → Social aspects of security and privacy; • **Social and professional topics** → **Governmental regulations**.

KEYWORDS

Auditing, Evaluation, Governance, Regulation, Policy, Risk, Fairness, Black-Box Access, White-Box Access, Adversarial Attacks, Interpretability, Explainability, Fine-Tuning

ACM Reference Format:

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, Dylan Hadfield-Menell, and . 2024. Black-Box Access is Insufficient for Rigorous AI Audits . In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 19 pages. <https://doi.org/3630106.3659037>

*Equal Contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

FAccT ’24, June 03–06, 2024, Rio de Janeiro, Brazil

© 2024 ACM.

ACM ISBN 979-8-4007-0450-5/24/06

<https://doi.org/3630106.3659037>

CONTENTS

Abstract	1
Contents	2
1 Introduction	2
2 Background	3
2.1 Black, Grey, White, and Outside-the-Box Access	3
2.2 Regulatory Frameworks' Reliance on Audits	3
2.3 Audits in the Status Quo	5
3 Limitations of Black-Box Access	5
4 Advantages of White-Box Access	6
4.1 White-box attack algorithms are more effective and efficient.	6
4.2 White-box interpretability tools aid in diagnostics.	6
4.3 Fine-tuning reveals risks from latent knowledge or post-deployment modifications.	7
5 Advantages of Outside-the-Box Access	7
6 Methods to Address Security Risks	8
7 Discussion	9
Acknowledgments	9
References	10
A Motivations for External Audits	18
B Technical Assistance as a form of Outside-the-Box Access	18
C Innovation on Auditing Tools	18
D Beyond Access: Other Aspects of Rigorous Audits.	19

1 INTRODUCTION

External evaluations of AI systems are emerging as a key component of AI oversight [10, 11, 19, 25, 32, 151, 160, 165, 177, 190–192, 194, 196, 201, 247, 249, 252, 272, 274, 276, 279, 284, 293, 294, 324, 333] and governance frameworks [61, 85, 217, 311, 314]. There is a rich history in academic AI research of evaluating systems to explain their behaviors [4, 343] and evaluate risks including those related to privacy [43, 285, 289, 291, 337], intellectual property rights [40, 47, 139], fairness and discrimination [12, 28, 37, 59, 64, 75, 81, 102, 154, 188, 302, 327], harmful content, [23, 24, 244, 253, 304, 305], circumvention of safeguards (*jailbreaks*) [42, 180, 255, 277, 281, 325, 355], misinformation and deception [125, 127, 133, 231, 271], dangerous capabilities [54, 56, 147, 222, 284], and broader societal impacts [294, 328].

Historically, most academic work on evaluating AI systems has been conducted on models where parameters, data, and methodology are openly available. AI systems that are not available to the public, including ones that are proprietary or in pre-deployment, pose challenges for oversight. AI audits are structured evaluations designed to identify risks and improve transparency by assessing how well models and methods meet specific desiderata [67, 202]. Norms for AI audits are not yet well established, and their effectiveness can vary depending on the degree of system access granted to auditors [11, 33, 252]. This is crucial because existing calls for audits are often agnostic to the form of access, and industry actors have previously lobbied for limiting access given to auditors.¹

¹For example, Google made the unsubstantiated claim in 2021 that “There will always be better methods for verifying the performance of an AI system (ex: input/output

auditing) than direct access to source code,” while lobbying the European Union on a draft of the EU AI Act [108].

Recently, some developers of prominent state-of-the-art AI systems have kept most details of their models private [29]. To public knowledge, voluntary external audits of these systems have primarily involved analysis of the input/output behavior of models [13, 192, 221, 306]. This form of access, in which auditors are only able to see outputs for given inputs, is known as *black-box*. Unfortunately, black-box access is very limiting for auditors. Some problems, such as anomalous failures, are difficult to find with black-box access [150], and others, such as dataset biases, can be actively reinforced by testing data [278].

The ability to query a black-box system is useful, but many of today’s evaluation techniques require access to weights, activations, gradients, or the ability to fine-tune the model [34]. *White-box* access refers to the unrestricted ability to observe a system’s internal workings. It enables evaluators to apply more powerful attacks to automatically identify weaknesses [107, 228], study internal mechanisms responsible for undesirable model behaviors [132, 153, 164], and identify harmful dormant capabilities through fine-tuning [242, 342]. Meanwhile, *outside-the-box* access involves additional contextual information about a system’s development or deployment such as methodology, code, documentation, hyperparameters, data, deployment details, and findings from internal evaluations. It allows auditors to study risks that stem from methodology or data [24, 46, 195, 278] and makes it easier to design useful tests. This paper makes four contributions:

- (1) We present shortcomings of black-box methods for evaluating AI systems (Section 3).
- (2) We overview the ways in which white-box methods involving attacks, model interpretability, and fine-tuning substantially expand the capabilities of evaluators (Section 4).
- (3) Similarly, we examine how outside-the-box access, including methodology, code, documentation, hyperparameters, data, deployment details, and findings from internal evaluations, allow for more thorough evaluations (Section 5).
- (4) Finally, we describe methods to conduct white- and outside-the-box audits securely to avoid leaks of sensitive information. These include *technical* solutions involving application programming interfaces, *physical* solutions involving secure research environments, and *legal* mechanisms that have precedent in other industries with audits (Section 6).

Given the growing evidence that different forms of access can facilitate very different levels of evaluation, we draw two conclusions. First, transparency regarding model access and evaluation methods is necessary to properly interpret the results of an AI audit. Second, white- and outside-the-box access allow for substantially more scrutiny than black-box access alone. When higher levels of scrutiny are desired, audits should be conducted with higher levels of access.²

²Jointly with this paper, we submitted a [comment on policy questions related to white- and outside-the-box audits](#) to a request for information from the US National Institute of Standards and Technology [207].

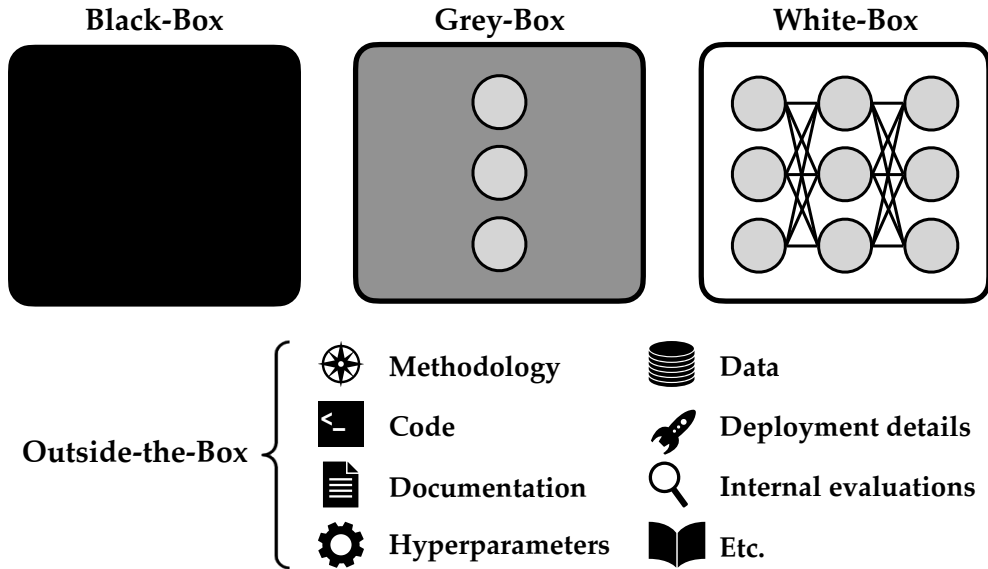


Figure 1: *Black-box* access lets auditors query the system and analyze the resulting outputs. *Grey-box* access lets auditors access limited internal information. *White-box* access lets users access the full system. *Outside-the-box* access gives auditors contextual information. In this paper, we argue that white- and outside-the-box access are key for rigorous AI audits.

2 BACKGROUND

2.1 Black, Grey, White, and Outside-the-Box Access

In accordance with literature on security and software testing [144], we differentiate black-, grey-, and white-box access. We also introduce two new concepts: “de facto white-box” access and “outside-the-box” access. Figure 1 illustrates these categories, and Table 1 summarizes which techniques each type of access allows.

- (1) **Black-box** access allows users to design inputs for a system, query it, and analyze the resulting outputs.
- (2) **Grey-box** access offers users limited access to a system’s inner workings. For neural networks, this can include information such as input embeddings, inner neuron activations, or sampling probabilities. There are many ways that users can be given information about a system’s inner workings and many corresponding shades of grey. **De facto white-box** access is a very light grey form of access that allows users to run arbitrary processes on a system indirectly with the constraint that the system’s parameters cannot be copied. We discuss this and other methods to minimize the possibility of leaks in Section 6.
- (3) **White-box** access allows users full access to the system. This includes access to weights, activations, gradients, and the ability to fine-tune the model.
- (4) **Outside-the-box** access grants users access to additional information about the system’s development and deployment. There are many types, which can include methodological details, source code, documentation, hyperparameters, training data, deployment details, and findings from internal evaluations. Different forms of outside-the-box access can vary

greatly in their comprehensiveness. For example, possessing high-level details (such as a “model card,” [195]) is less informative compared to having comprehensive documentation from training and testing.

2.2 Regulatory Frameworks’ Reliance on Audits

Emerging frameworks for AI governance have been designed to rely on high-quality audits. Audits have been called for in the White House Executive Order on AI [217], European Union policy [85–87], other policy initiatives [61, 68, 211, 311], general AI principles [60, 94, 208, 216, 312], voluntary standards [214, 235, 314], multilateral commitments [5], and position papers [209, 236]. In particular, audits in these proposals are intended to provide trustworthy *assessments* of potential harm and *explanations* of system behaviors.

Regulatory frameworks have called for evaluations to accurately assess risks. Some jurisdictions may require risk assessment evaluations for AI systems used in certain contexts. These can include tests to ensure non-discrimination, such as New York City’s requirement for bias audits of automated employment decision tools [68], or quality and performance evaluations [116]. The draft EU AI Act [87] has more recently harmonized quality assurance standards across several high-risk use cases, with provisions for external oversight. Regulators are also increasingly interested in external oversight of AI systems with potentially harmful capabilities. Recently, U.S. Executive Order 14110 [217] required developers of certain foundation models to share test results with the Federal Government. It also instructed the National Institute of Standards and Technology (NIST) to develop evaluation guidelines for harmful AI capabilities, and it tasked the Department of Energy with developing tools and testbeds to evaluate threats from AI systems to security and critical infrastructure. Companies may also voluntarily

	Access Level	Black-Box	Grey-Box	De facto White-box	White-Box	Outside-the-box
Test sets (Section 3)	Queries	✓	✓	✓	✓	✗
Manual attacks (Section 3)		✓	✓	✓	✓	✗
Transfer-based attacks (Section 4.1)		✓	✓	✓	✓	✗
Gradient-free attacks (Section 4.1)		✓	✓	✓	✓	✗
Sampling-probability-guided attacks (Section 4.1)	Probabilities	✗	✓	✓	✓	✗
Gradient-based attacks (Section 4.1)	Gradients	✗	✗	✓	✓	✗
Hybrid attacks (Section 4.1)		✗	✗	✓	✓	✗
Latent space attacks (Section 4.1)	Weights/ Activations	✗	✗	✓	✓	✗
Mechanistic interpretability (Section 4.2)		✗	✗	✓	✓	✗
Fine-tuning (Section 4.3)	Fine-tuning	✗	✗	✓	✓	✗
Methodological evaluations (Section 5)	Outside-the-Box	✗	✗	✗	✗	✓
Data evaluations (Section 5)		✗	✗	✗	✗	✓
Complementary evaluations (Section 5)		✗	✗	✗	✗	✓
Using source code (Section 5)		✗	✗	✗	✗	✓
Copying system parameters (Section 6)	Unrestricted	✗	✗	✗	✓	✗

Table 1: A summary of what evaluation techniques are possible with which types of access. A ✓ means that a technique is possible while an ✗ means it is not. Many levels of grey-box access are possible, but we highlight sampling-probability-attacks because they are a common example.

Goal	Technique	Advantage
Identifying problems	Attacks	White-box methods are more reliable for detecting anomalous failures.
		White-box attacks are more efficient.
		White-box methods provide auditors with more attack options.
		Robustness to white-box attacks confers greater robustness assurances.
	Fine-tuning	Fine-tuning enables searching for harmful dormant capabilities.
Interpretability	Interpretability tools enable stronger assurances of what knowledge and capabilities a system has.	
	Outside-the-box assessment	Information helps auditors target complementary evaluations.
Incentivizing responsible development	Outside-the-box assessment	Information about deployment assists with assessing societal impacts.
		Auditors can trace problems to datasets and methodological decisions made by developers.
		Auditors can assess risk-mitigation strategies.
Increasing transparency	Interpretability	White-box methods can address misconceptions about model outputs, characteristics, and risks.
		White-box methods assist with providing more accurate and meaningful explanations.
Enabling debugging	Attacks	Stronger attacks produce stronger instances of failures to incorporate into training data.
	Interpretability	Interpretability techniques enable more precise debugging.

Table 2: A summary of the advantages that various white-box and outside-the-box auditing techniques provide over black-box methods that are discussed in Section 4 and Section 5. In Appendix A, we expand on the variety of motivations for AI audits.

subject their systems to external evaluations beyond regulatory requirements. For example, the NIST Risk Management Framework provides recommendations for audits related to system design and reliable operation [314]. In Section 3, Section 4, and Section 5, we overview advantages of white- and outside-the-box access over black-box access for rigorous assessments.

Assessing model explanations enables scrutiny of automated decisions. Regulatory frameworks also use audits to provide those affected by automated decision-making with explanations of the decisions [111, 238]. In some jurisdictions, such as the European Union, when individuals are harmed by automated decision-making systems, they may have the right to an explanation, and the results of these explanations may entitle them to

remediation [36, 116]. Further, explanation requirements may exist for particularly high-risk systems, such as EU platform regulations (the Digital Markets Act and the Digital Services Act) that require transparency from large online platforms using AI systems (e.g., ranking algorithms) to protect against discrimination and abuse of market power [88, 115]. Finally, disclosure of evidence may be required under liability rules, such as the Product and AI Liability Directives in the EU, to enable potential claimants to adequately defend damage claims [35, 114]. When producing explanations, a report from NIST emphasizes the importance of explanation accuracy, or *[a]n explanation [that] correctly reflects the reason for generating the output and/or accurately reflects the system’s process* [237]. In Section 3 and Section 4.2, we overview advantages of white-box access over black-box access for generating reliable explanations.

2.3 Audits in the Status Quo

Recent advancements in AI capabilities – especially from large generative models – have increased public attention on AI audits. As of January 2024, there are no widely adopted norms for conducting AI audits [25]. Details of AI audits can vary because they depend upon the system, how it will be used, and what risks it poses. Auditing frameworks and metrics have been proposed for specific use cases, including hiring [142, 245], facial recognition [143, 250], healthcare [177, 183], recommender systems [57, 260], and general purpose language models [202]. However, Raji et al. [252] identifies five *general* limitations for algorithmic audits: scope, independence, level of access, professionalism, and public disclosure of methods and results.

Currently, evaluations of proprietary or pre-deployment AI systems are predominantly performed in-house by developers with selective disclosure of methods and outcomes. Some developers have voluntarily partnered with external auditors and provided them with black-box access to state-of-the-art systems [13, 147, 192, 222, 306]. Additionally, some developers run programs for external researchers to support their internal evaluation process (e.g., OpenAI’s Preparedness Challenge [223] and Red-Teaming Network [224]). However, to public knowledge, these industry-based efforts have involved black-box and limited outside-the-box access such as *model cards* [195].

3 LIMITATIONS OF BLACK-BOX ACCESS

Black-box evaluations of AI systems are based on analysis of their inputs and outputs only. Such evaluations often involve assessing performance on test sets [121, 172, 269, 297, 299, 320] or searching for inputs that elicit harmful outputs [51, 147, 233, 234, 271, 325]. Generative AI audits often attempt to elicit undesirable capabilities or behaviors (e.g., [30, 118, 140, 147, 198, 206, 266, 284, 292, 301]). However, black-box methods are inherently limited in their ability to identify harms or provide meaningful explanations. Readers with a computer science background can consider the analogy of attempting to evaluate the performance of software without reading or modifying its source code.

Black-box methods are not well suited to develop a generalizable understanding. Black-box access limits evaluators to analyzing a system using only inputs and outputs. However, the vast number of possible inputs to AI systems makes it intractable

to develop a complete understanding from this alone. This forces evaluators to rely on heuristics to produce ‘relevant’ inputs for evaluation. For this reason, black-box methods have been shown to be unreliable for detecting failures that elude typical test sets including jailbreaks, adversarial inputs, or backdoors [58, 325, 353].

Black-box access prevents system components from being studied separately. Analyzing components of a system separately is ubiquitous in science and engineering. It enables engineers to trace problems to support more targeted interventions. However, black-box access obscures what subsystems the AI system is composed of. For example, black-box access does not allow input or output filters to be studied separately from the rest of the system. Other issues can arise from a lack of outside-the-box access to data. Datasets can inform evaluations related to privacy and copyright [262], and can help to avoid problems from data contamination [72, 105, 131]. Having a lack of outside-the-box knowledge about how the system is deployed also prevents evaluators from making a more practical assessment of broader societal impacts [328].

Black-box evaluations can produce misleading results. Since black-box evaluations rely entirely on the queries made to the system, they are biased by how evaluators design inputs [138, 205, 326]. This can lead to misconstrued conclusions about the system’s characteristics. For example, systems may satisfy simple statistical tests for non-discrimination, but they may still have undesirable biases in their underlying reasoning [246]. Developers with information about the black-box tests can exacerbate this problem by modifying the model’s output behavior on test cases despite unresolved flaws in its internal reasoning. In addition, Schaeffer et al. [270] provide examples of black-box prompt-based evaluation methods for language models that can lead to misunderstandings of their emergent capabilities.

Black-box explanation methods are often unreliable. Using black-box methods alone to produce explanations for an AI system’s decisions is difficult [117, 264]. Many black-box techniques to provide counterfactual explanations for model decisions are misleading because they fail to reliably identify causal relationships between the system’s input features and outputs [62]. Explanation methods for black-box systems can also be exploited by adversaries to produce misleading explanations for harmful decisions [6, 290]. Furthermore, when generative language models are asked to explain their decisions, their justifications do not tend to be faithful to their actual reasoning [310].

Black-box evaluations offer limited insights to help address failures. Black-box evaluations offer little insight into ways to address problems they discover. The main technique they enable is to train on problematic examples, but this can fail to address the underlying problem [78, 97], be sample-inefficient [322], and may introduce new issues. Corrective actions are not robust when they fail to address a problem at its root. For example, some recent works have shown that safety measures built into large language models can be almost entirely undone by fine-tuning on a small number of harmful examples [168, 242, 336, 342]. In contrast, white-box methods reveal more about the nature of flaws, facilitating more precise debugging methods [322].

4 ADVANTAGES OF WHITE-BOX ACCESS

White-box offers a wider range of techniques to detect symptoms, understand causes, and mitigate harms in a targeted manner [34]. Even for a system that will only be deployed as a black box, white-box audits are still more useful for finding problems. Here, we survey techniques for white-box evaluations and their advantages over black-box ones.

4.1 White-box attack algorithms are more effective and efficient.

In machine learning, *adversarial attacks* refer to inputs that are designed specifically to make a system fail. AI systems have a long history of having unexpected failure modes that can be triggered by very subtle features in their inputs [7, 346]. Attacks play a central role in evaluations because they help to assess a system’s *worst-case behavior*.

White-box algorithms produce stronger attacks. White-box algorithms allow for gradient-based optimization of adversarial inputs, which is powerful compared to simpler search methods. For example, white-box adversarial attack algorithms against vision systems typically use the gradient of the adversarial objective with respect to the input pixels to design adversarial inputs [107, 228]. This is much more effective for finding vulnerabilities than unguided black-box search methods. Consequently, white-box attacks are dominant in vision applications. In reinforcement learning, white-box access to a target agent also helps develop stronger adversarial attacks against it [49, 323]. For language models, optimizing adversarial inputs with gradient-based methods is more challenging because text (unlike pixels) is discrete which prevents gradient propagation. Nonetheless, there are various state-of-the-art white-box techniques for attacking language models. These include using a differentiable approximation to the process of sampling text [112, 295, 318], projecting adversarial embeddings onto text embeddings [329], and performing gradient-informed searches over modifications to textual changes [80, 136, 170, 175, 258, 287, 355].

Many black-box and grey-box attack algorithms are simply indirect or inefficient versions of white-box ones. Many black-box attacks against AI systems involve attacking a white-box model with a white-box algorithm and then testing the resulting attack on the target black-box model [179, 350]. For vision models, the main motivation behind studying black-box attacks is that white-box access is not always available to attackers [21]. Additionally, several of the most effective attacks against state-of-the-art models, such as GPT-4 and Claude-2, have simply been the result of transferring a white-box attack generated against an open-source model to the intended black-box target model [355]. Other types of black- and grey-box attack algorithms involve inefficiently estimating gradients by analyzing outputs or sampling probabilities across many queries [110, 128] when more precise gradients could be obtained trivially with white-box access.

Latent space attacks help to make stronger assurances. Typically, AI systems are attacked by crafting *inputs* meant to make them exhibit undesirable behavior. However, input space attacks are not well-suited to diagnose certain hard-to-find issues, including

high-level misconceptions [300], anomalous failures [353], backdoors [58, 334], and deception [63, 231]. A complementary technique for attacking systems in the input space is to relax the problem and attack their internal *latent representations*. The motivation of latent space attacks is that some failure modes are easier to find in the latent space than in the input space [53] because concepts important to the system’s reasoning are represented at a higher level of abstraction inside the model [16, 135, 275, 309, 354]. Thus, robustness to latent attacks enables evaluators to make stronger assurances of safe worst-case performance. Latent space attacks are also more efficient to produce because they require less gradient propagation than input space attacks [229, 243], allowing for more thorough debugging work to be conducted on a limited time and computing budget. Latent space attacks are still an active area of research, but some works have emerged showing that robustness to latent space attacks effectively indicates robustness to input space attacks in vision models [158, 178, 226, 267, 344, 351]. Since textual inputs to language models are discrete, only latent space attacks allow for the direct use of gradient-based optimization, rendering them especially useful for language models [119, 134, 148, 156, 171, 176, 227, 265, 352].

White-box methods expand the attack toolbox. Some black-box attack methods are competitive, particularly against language models. These include methods based on local search [240], rejection sampling at scale [96], Langevin dynamics [157, 286], evolutionary algorithms [162], and reinforcement learning [51, 74, 233]. Additionally, some of the most effective methods for attacking language models involve human or human-guided generation of adversarial prompts [180, 277, 325]. However, even when black-box attacks are useful, white-box algorithms are complementary because they generate qualitatively different kinds of attacks. For example, many black-box techniques produce attacks that appear as natural language (e.g., [277]) while white-box algorithms are state-of-the-art for synthesizing adversarial prompts which appear to humans as unintelligible text (e.g., [355]). Relying only on black-box methods would likely miss out on these types of failures entirely because finding them with black-box searches or human inductive biases would be highly improbable. Black- and white-box methods can also be combined to conduct *hybrid* attacks by using the results of one method as an initialization for another. Combinations of attacks tend to be better at helping humans find vulnerabilities than a single method alone [50].

4.2 White-box interpretability tools aid in diagnostics.

While it is possible to infer properties of a system from studying inputs and outputs, understanding its internal processes allows evaluators to more thoroughly assess its trustworthiness [103, 256]. Interpreting the inner mechanisms of models has been recognized as a key part of agendas for reducing harms from AI systems [126, 137, 212], and explaining how models make specific decisions has also been recognized as a way to protect the rights of individuals affected by AI [111, 238].

White-box interpretability tools help evaluators discover novel failure modes. White-box algorithms and interpretability tools have aided researchers in finding vulnerabilities. Examples

have involved identifying novel attacks [353], internal representations of spurious features [95, 184], brittle feature representations [44, 48, 50, 52, 101, 123, 199, 335, 341], and limitations of key-value memories in transformers [99, 100, 189]. As an added benefit, attributing the problem to specific parts of the system’s architecture or representations allows developers to address it in a more precise way [322].

Studying internal representations can help to establish the presence or lack of specific capabilities. White-box methods allow for more precise identification of what knowledge and capabilities a system has [8, 18]. Tools such as concept vectors [2, 146] and probes [66] allow humans to assess the extent to which system internals can be understood in terms of familiar concepts. For example, these techniques have been used to study features related to fairness in visual classifiers [109], provide evidence that language models internally represent space and time [113], and show that networks sometimes represent truth-like features along linear directions [38, 185]. Methods for this are imperfect and still an active area of research [14, 84, 257], but interpretations like these offer a potentially powerful way to identify whether a model represents specific concepts.

Consider an example. Suppose that an auditor wants to assess *sycophancy*: a language model’s tendency to pander to the biases of users who chat with it [234, 280]. For example, an evaluator might be concerned that the system will respond differently when the user says they are conservative or liberal in the chat. Black-box techniques could only be used to argue that the system is not sycophantic by producing examples and analyzing them for apparent sycophancy. However, a white-box interpretability-based approach could offer much more information. For example, if it were not possible for a classifier to distinguish whether the user revealed themselves to be conservative or liberal from the model’s internal representations, then this would offer stronger evidence that the system will reliably not exhibit this type of sycophancy.

Mechanistic understanding helps to make stronger assurances. In general, it is impossible to make guarantees about black-box systems using a finite number of queries without additional assumptions. In contrast to black-box methods, which can only show the existence of failures by finding inputs that elicit them, thoroughly understanding the computations inside of a model gives auditors a complementary way to find evidence against the existence of failure modes. A mechanistic understanding can help researchers develop a predictive model of how the system would act for broad classes of inputs. Some works have aimed to provide thorough investigations of how networks perform simple tasks [39, 204, 348]. Although scaling thorough analysis is an open challenge [303], it offers a strategy for making strong assurances. Recent works have attempted to make progress on this problem by using sparse autoencoders to allow evaluators to more thoroughly study the features represented inside of large language models [31, 69, 184].

White-box methods expand the toolbox for explaining specific AI system decisions. As discussed in Section 2, existing regulatory frameworks have been designed with specific desiderata for model explanations in order to determine accountability and protect individual rights. Many techniques are used to provide explanations of model behaviors during audits [343]. Black-box techniques can only attribute decisions to input features using techniques that

involve modifying inputs and analyzing how model outputs change [259]. However, these techniques are frequently misleading [264] and can fail to reliably identify causal relationships between the system’s input features and output [62]. White-box access expands and strengthens the toolbox by allowing for gradient-based techniques [70, 79, 173, 347]. It also allows for explainability tools to be combined with interpretations of the model mechanisms to explain a model’s behaviors in terms of more abstract concepts.

4.3 Fine-tuning reveals risks from latent knowledge or post-deployment modifications.

State-of-the-art AI systems are typically trained on large amounts of internet data, often in multiple stages. This can cause them to learn undesirable capabilities, such as knowledge of how to perform illegal activities [277, 354] or the ability to produce harmful content [23, 24, 98, 182, 253, 278, 304, 305]. Developers attempt to remove harmful abilities through fine-tuning, but they can unexpectedly resurface through “jailbreaks” [9, 55, 73, 169, 180, 220, 241, 255, 277, 282, 308, 325, 339, 340, 355] or further fine-tuning models on a small number of new examples [168, 242, 336, 342]. The existence of harmful dormant capabilities in models thus poses risks from attacks and fine-tuning, especially if they are leaked (e.g. Stable Diffusion [161]), open-sourced (e.g., Llama-2 [306]), or deployed with fine-tuning access via API (e.g., GPT-3.5 [221]). Consequently, being able to fine-tune the model offers another strategy to search for evidence of undesirable capabilities and assess the risks in deployment.

5 ADVANTAGES OF OUTSIDE-THE-BOX ACCESS

In addition to having access to AI systems themselves, giving auditors outside-the-box access to contextual information also helps to identify risks. This can include methodological details, source code, documentation, hyperparameters, training data, deployment details, and the findings of internal evaluations. While it can come in many types, all outside-the-box information can be useful to auditors for three common reasons: (1) helping auditors more effectively design and implement tests, (2) offering clues about potential issues, and (3) helping auditors trace problems to their sources. See also Appendix B where we discuss how outside-the-box access to technical assistance from developers can also be useful for auditors.

Code, documentation, and hyperparameters help auditors work more efficiently. As discussed in Section 4, audits can require a number of technical evaluations. Having code and documentation from developers can streamline the process of designing them. For example, consider fine-tuning evaluations. Fine-tuning a model typically requires precisely configured code and hyperparameters that have been carefully selected after extensive testing, often over the course of weeks or months. Using the developer’s existing resources is a much more efficient option for auditors compared to re-implementing everything from scratch.

Access to methodological details helps to identify risks. Knowing methodological details can reveal shortcuts taken during development, which can guide evaluators toward discovering problems. For example, if a system was trained with human-generated

data using a non-representative cohort of humans, this can suggest specific social biases that the system may have internalized [46, 268, 298]. Knowing the findings of internal evaluations is especially useful for helping auditors target their efforts toward a set of complementary evaluations. Furthermore, when developers attempt to mitigate flaws, auditors can better assess the effectiveness of these efforts if they have detailed information about the attempted mitigation (e.g., fine-tuning datasets, both old and new versions of model weights, etc.) [239].

Access to data helps auditors trace problems and assess fair use. Recent work has highlighted the ability of dataset audits to identify harmful and biased content used to train models [23, 24, 98, 182, 278, 304, 305]. For example, Birhane et al. [24] and Thiel [304] were able to identify previously-overlooked examples of unintended hateful, sexual, and child-abuse-related content in widely-used training datasets. Access to training data also helps to investigate risks of data-poisoning attacks (which is especially important for systems trained on internet data) [41, 254, 319, 321]. Meanwhile, legal questions are currently being debated involving the extent to which training generative AI systems on copyrighted content constitutes fair use [65, 120, 139, 261]. Auditors may require access to training data to properly assess whether it was used in accordance with copyright law.

Contextual information makes it easier to hold developers accountable. Requirements to produce documentation place greater responsibility on developers to detail their methods, especially if subject to regulatory penalty defaults [332, 338]. Contextual information provides information about whether developers made decisions in a responsible manner [200]. For example, documentation can provide insights into why certain design choices were made over others [190]. Datasets and training details can help trace risks to intentional choices, and internal evaluation reports provide insights into how the developer responded to findings. By increasing the scrutiny placed on decisions in the development process, requirements for greater methodological transparency to auditors can deter developers from taking risks in the first place [46, 159].

6 METHODS TO ADDRESS SECURITY RISKS

A concern with white- and outside-the-box audits is an increased risk that a developer’s models or intellectual property could be leaked [29]. In turn, leaks could compromise developers’ trade secrets and pose risks to the public if they enable misuse [210]. Widespread norms for secure audits in AI do not yet exist, but there is precedent in other fields for navigating similar challenges to enable secure oversight. The risk of leaks can be minimized through several *technical*, *physical*, and *legal* mechanisms. With these measures, developers can provide white- and outside-the-box access to auditors without the system’s parameters leaving their servers. These can reduce leakage risks to a level comparable to ones posed by common existing practices.

Technical: API access can offer remote auditors de facto white-box access. Forms of structured access, particularly research application programming interfaces (APIs) [26, 34, 92, 283], could enable auditors to analyze systems using some white-box tools without giving auditors direct access to model parameters. We refer to this form of access as *de facto* white-box access if it enables

auditors to indirectly run arbitrary white-box processes on models while restricting direct access to model parameters. One example of running an algorithm that accesses a model’s parameters via API is an OpenAI GPT-3.5 API which allows for fine-tuning [221]. However, more customizable APIs (e.g., [92]) would be needed to allow for more flexible access. Another proposed paradigm is a *flexible query API* [225], where auditors are given complete access to mock versions of a model and data. The auditors then develop evaluations using their complete access to these mock artifacts before submitting them to be run on the true model and dataset. This allows auditors to better customize their evaluations.

The goal of API access is to ensure that the system cannot easily be reconstructed. However, prohibiting the sharing of weights is neither necessary nor sufficient for this. For example, sharing a small subset of weights with auditors is unlikely to pose significant security risks [187], but sharing other information, such as the product of weights with their pre-synaptic neuron’s activations, may allow for parameter reconstruction. This suggests the need for a process by which a developer can raise grievances about specific requests from auditors and have them adjudicated. Overall, while conceptually simple, designing APIs that simultaneously provide the comprehensiveness, flexibility, and security required for rigorous auditing is an open area of research [34]. Greater clarity is required regarding how to balance different desiderata. For example, more comprehensive access may impact security by facilitating model reconstruction, as discussed above. In Appendix C, we also discuss how investment, research, and development into secure auditing infrastructure can help with progress toward improved techniques.

Physical: Secure research environments can be used for auditors given unrestricted white-box access. Auditing personnel could securely be given white-box access to a system by hosting them on-site at the developer’s facilities in a secure research environment. This is a common practice in other industries despite the costs of requiring auditors to be physically on-site [225, 307, 315]. For example, the International Atomic Energy Agency employs over 300 expert inspectors from approximately 80 countries who do on-site inspections of nuclear facilities [129, 130]. Compared to API access, secure research environments could allow auditors to access systems more flexibly and efficiently while minimizing the risk that the model is leaked or reconstructed. Safeguards for limiting information leakage through lab employees (such as NDAs) are already common practice and could be adapted for application to on-site auditors. However, it is unclear whether, absent external legal structures, labs could incentivize adherence to protective measures to the same extent as with employees.

Legal: Other industries have developed practices to address the risk of leaks from audits. Across many industries, auditors require privileged access to systems and data in order to perform effective assessments. There are established mechanisms from other fields, such as financial auditing, employed to reduce the risk of leaks. In the finance industry, this manifests in three main ways. First, policies for confidentiality and handling of sensitive information are enforced by formal training and non-disclosure clauses in contracts to hold auditors accountable for violations [89]. Second, there are clear terms of engagement that govern the relationship

between auditor and auditee. These typically include specific restrictions on confidentiality expectations tailored to a particular client [15]. While some specifics of auditing are managed through contracting, the Public Company Accounting Oversight Board (PCAOB) requires all registered auditors to adhere to common standards in the US [232]. Finally, auditors can be legally required to avoid conflicts of interest. In the US, this is done through a regime specifying general provisions for auditor independence (such as reporting requirements) outlined in Title II of the Sarbanes Oxley Act [232]. Provisions are enforced through various agencies, including the Securities Exchange Commission (SEC), which prevents auditor manipulation or the use of financial information for personal gain [215]. This type of enforcement could allow for auditors to be held accountable in a way that could reduce the risk of leakage to a level comparable to risks posed by the developer’s employees. In fact, employees may pose greater risks of sharing tacit knowledge with competitors than auditors because developers regularly attempt to recruit AI researchers from competitors.

7 DISCUSSION

White- and outside-the-box audits offer several benefits to developers. One potential benefit of more rigorous audits for developers is increased credibility by creating the perception that their systems are of higher quality. Meanwhile, white- and outside-the-box evaluations also offer developers greater insight into addressing problems with systems they build [248]. We discuss in Section 3 how black-box evaluations can only establish when problems exist, while white- and outside-the-box methods can provide a clearer diagnosis to help address them. For example, if a flaw in a system can be attributed to a specific set of components, this enables more targeted interventions to fix it.

However, absent legal requirements, developers have strong incentives to limit access granted to external auditors. Thus far, external audits of state-of-the-art AI systems, when they have occurred, have been black-box (to public knowledge) [13, 147, 192, 221, 306], suggesting a failure of existing incentive structures to provide greater access to auditors. Developers are typically reluctant to provide more permissive access to their models and related resources [252]. This may stem from concerns that information collected through white- and outside-the-box audits could be leaked [29] (though this is addressable—see Section 6). Furthermore, it could expose a system’s lackluster performance, vulnerabilities, or poor risk management processes by developers, which could lead to reputational harm and possibly legal liability [166].

Current black-box audits may set a precedent for future norms. Established norms frequently become “sticky” and entrenched in regulatory regimes [213]. Accordingly, the current norm for black-box audits [13, 147, 192, 221, 306] may set the future standard. Furthermore, in policy debates about audits, industry actors have also lobbied for limiting external auditors to black-box access [108]. Without sufficient access and resources, non-industry researchers will struggle to iterate upon methods for more thorough audits [34]. Over time, this could limit or bias public understanding of AI systems. A lack of open research on transparency tools and the view that there is little social benefit from greater transparency are mutually reinforcing. Appendix C expands on how

investment, research, and development into auditing techniques and infrastructure can facilitate further progress.

Low-quality audits can be counterproductive. Poor (e.g., black-box) audits can have counterproductive effects: they can increase public or regulatory trust in systems on false grounds, preventing appropriate levels of external scrutiny [93, 174]. They also enable safety- or ethics-washing by developers [90, 152, 330] who make AI systems that contribute to risks without sufficiently investing in methods to address them.

Conclusion: We have argued that providing auditors with white-box and thorough outside-the-box access to systems is feasible and allows for more meaningful oversight from audits. We draw two conclusions. First transparency regarding model access and evaluation methods is necessary to properly interpret the results of an AI audit. Second, white- and outside-the-box access allow for substantially more scrutiny than black-box access alone. When higher levels of scrutiny are desired, audits should be conducted with higher levels of access. Finally, we emphasize that white-box and thorough outside-the-box access are necessary but not sufficient for rigor. Audits can and do fail for many reasons. Without careful institutional design, the incentives of developers and auditors may result in audits that do not consistently align with public interest [11, 25, 67, 218, 251, 252]. In Appendix D, we examine additional ways in which the quality of audits can be compromised.

ETHICS STATEMENT

Given the role of rigorous audits in improving accountability and representing public interests, we expect the foreseeable impacts of this paper to be positive. However, audits can fail to benefit the public for a variety of reasons. We discuss these challenges in Appendix D.

CONTRIBUTIONS

Carson Ezell and Stephen Casper were the central writers and organizers. Charlotte Siegmann, Kevin Wei, Andreas Haupt, and Taylor Curtis contributed primarily to Section 2. Noam Kolt contributed primarily to Section 7 and Appendix D. Jérémy Scheurer, Marius Hobbhahn, and Lee Sharkey contributed primarily to Section 3, Section 4.2, Section 7, Appendix C, and Appendix D. Satyapriya Krishna contributed primarily to Section 4.1, Marvin von Hagen and Silas Alberti contributed primarily to Appendix A and Section 2.3. Qinyi Sun, Michael Gerovitch, and Benjamin Bucknall contributed primarily to Section 2.1, Section 6, and Section 7. Alan Chan, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell offered high-level feedback and guidance.

ACKNOWLEDGMENTS

We are grateful for discussions and feedback from Markus Anderljung, Thomas Krendl Gilbert, Matthijs Maas, Javier Rando, Stewart Slocum, Luke Bailey, Adam Jermyn, Alexandra Bates, Miles Wang, Audrey Chang, and Erik Jenner.

REFERENCES

- [1] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 287–297.
- [2] Abubakar Abid, Mert Yuksekgonul, and James Zou. 2022. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*. PMLR, 66–88.
- [3] Julius Adebayo, Michael Muell, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429* (2020).
- [4] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems* 35 (2022), 15784–15799.
- [5] AI Safety Summit. 2023. The Bletchley Declaration by Countries Attending the AI Safety Summit. <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>
- [6] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 161–170. <https://proceedings.mlr.press/v97/aivodji19a.html> ISSN: 2640-3498.
- [7] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access* 6 (2018), 14410–14430.
- [8] Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. (2018). [arXiv:1610.01644](https://arxiv.org/abs/1610.01644) [stat.ML]
- [9] Alex Albert. 2023. Jailbreak Chat. (2023). <https://www.jailbreakchat.com/>
- [10] Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718* (2023).
- [11] Markus Anderljung, Everett Thornton Smith, Joe O’Brien, Lisa Soder, Benjamin Bucknall, Emma Blumke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. 2023. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. (2023). [arXiv:2311.14711](https://arxiv.org/abs/2311.14711) [cs.CY]
- [12] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.
- [13] Anthropic. 2023. Challenges in evaluating AI systems. (2023). <https://www.anthropic.com/index/evaluating-ai-systems>
- [14] Omer Antverg and Yonatan Belinkov. 2021. On the pitfalls of analyzing individual neurons in language models. *arXiv preprint arXiv:2110.07483* (2021).
- [15] Compiled Auditing Standard ASA. 2006. Auditing standard ASA 210 terms of audit engagements.
- [16] Ben Athiwaratkun and Keegan Kang. 2015. Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313* (2015).
- [17] Ramin P. Baghai and Bo Becker. 2020. Reputations and credit ratings: Evidence from commercial mortgage-backed securities. *Journal of Financial Economics* 135, 2 (Feb. 2020), 425–444. <https://doi.org/10.1016/j.jfineco.2019.06.001>
- [18] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48, 1 (2022), 207–219.
- [19] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. [n. d.]. Managing AI Risks in an Era of Rapid Progress. ([n. d.]).
- [20] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. 2023. Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688* (2023).
- [21] Siddhant Bhamri, Sumanu Muku, Avinash Tulasi, and Arun Balaji Buduru. 2019. A survey of black-box adversarial attacks on computer vision models. *arXiv preprint arXiv:1912.01667* (2019).
- [22] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.
- [23] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. *arXiv preprint arXiv:2311.03449* (2023).
- [24] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).
- [25] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. [arXiv:2401.14462](https://arxiv.org/abs/2401.14462) [cs.CY]
- [26] Emma Blumke, Tatum Collins, Ben Garfinkel, and Andrew Trask. 2023. Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases. (March 2023). <https://arxiv.org/abs/2303.08956v2>
- [27] Patrick Bolton, Xavier Freixas, and Joel Shapiro. 2012. The Credit Ratings Game. *The Journal of Finance* 67, 1 (2012), 85–111. <https://doi.org/10.1111/j.1540-6261.2011.01708.x> eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2011.01708.x>.
- [28] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
- [29] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. (Oct. 2023). <http://arxiv.org/abs/2310.12941> [arXiv:2310.12941](https://arxiv.org/abs/2310.12941) [cs].
- [30] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* (2023).
- [31] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholaschiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread* (2023). <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- [32] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (2021), 2053951720983865.
- [33] 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 Blumke, Jonathan Lebensold, Cullen O’Keefe, Mark Koren, Théo Ryyffel, 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 Szath, 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. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (April 2020). <https://doi.org/10.48550/arXiv.2004.07213> [arXiv:2004.07213](https://arxiv.org/abs/2004.07213) [cs].
- [34] Benjamin S Bucknall and Robert F Trager. 2023. Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers’ Model Access Requirements. (Oct. 2023). <https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/>
- [35] Miriam Buiten, Alexandre de Streeel, and Martin Peitz. 2021. EU Liability Rules for the Age of Artificial Intelligence. (April 2021). <https://doi.org/10.2139/ssrn.3817520>
- [36] Miriam C. Buiten, Louise A. Dennis, and Maïke Schwammberger. 2023. A Vision on What Explanations of Autonomous Systems are of Interest to Lawyers. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, Hannover, Germany, 332–336. <https://doi.org/10.1109/REW57809.2023.00062>
- [37] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [38] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827* (2022).
- [39] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill* (2020). <https://doi.org/10.23915/distill.00024> <https://distill.pub/2020/circuits>
- [40] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).
- [41] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149* (2023).
- [42] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, et al. 2023. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447* (2023).
- [43] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX*

- Security Symposium (USENIX Security 21)*. 2633–2650.
- [44] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Exploring neural networks with activation atlases. *Distill*. (2019).
- [45] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
- [46] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. (Sept. 2023). <https://doi.org/10.48550/arXiv.2307.15217> arXiv:2307.15217 [cs].
- [47] Stephen Casper, Zifan Guo, Shreya Mogulothu, Zachary Marinov, Chinmay Deshpande, Rui-Jie Yew, Zheng Dai, and Dylan Hadfield-Menell. 2023. Measuring the Success of Diffusion Models at Imitating Human Artists. *arXiv preprint arXiv:2307.04028* (2023).
- [48] Stephen Casper, Kaivalya Hariharan, and Dylan Hadfield-Menell. 2022. Diagnostics for deep neural networks with automated copy/paste attacks. In *NeurIPS ML Safety Workshop*.
- [49] Stephen Casper, Taylor Killian, Gabriel Kreiman, and Dylan Hadfield-Menell. 2023. Red Teaming with Mind Reading: White-Box Adversarial Policies Against RL Agents. (Oct. 2023). <http://arxiv.org/abs/2209.02167> arXiv:2209.02167 [cs].
- [50] Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. 2023. Red Teaming Deep Neural Networks with Feature Synthesis Tools. (Sept. 2023). <http://arxiv.org/abs/2302.10894> arXiv:2302.10894 [cs].
- [51] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv preprint arXiv:2306.09442* (2023).
- [52] Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. 2022. Robust feature-level adversaries are interpretability tools. *Advances in Neural Information Processing Systems* 35 (2022), 33093–33106.
- [53] Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. 2024. Defending Against Unforeseen Failure Modes with Latent Adversarial Training. arXiv:2403.05030 [cs.CR]
- [54] 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. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. 651–666. <https://doi.org/10.1145/3593013.3594033> arXiv:2302.10329 [cs].
- [55] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking Black Box Large Language Models in Twenty Queries. *arXiv preprint arXiv:2310.08419* (2023).
- [56] PV Charan, Hrushikesh Chunduri, P Mohan Anand, and Sandeep K Shukla. 2023. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads. *arXiv preprint arXiv:2305.15336* (2023).
- [57] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Transactions on Information Systems* 41, 3 (Feb. 2023), 67:1–67:39. <https://doi.org/10.1145/3564284>
- [58] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).
- [59] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. 2022. Fairness testing: A comprehensive survey and analysis of trends. *arXiv preprint arXiv:2207.10223* (2022).
- [60] China Academy of Information and Communications Technology and JD Explore Academy. 2021. White Paper on Trustworthy Artificial Intelligence. <https://cset.georgetown.edu/publication/white-paper-on-trustworthy-artificial-intelligence/>
- [61] Chinese National Information Security Standardization Technical Committee. 2023. Translation: Basic Safety Requirements for Generative Artificial Intelligence Services (Draft for Feedback). https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai/?utm_source=stack&utm_medium=email
- [62] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2021. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. (June 2021). <https://doi.org/10.48550/arXiv.2103.04244> arXiv:2103.04244 [cs].
- [63] Paul Christiano. 2019. Worst-case guarantees. <https://ai-alignment.com/training-robust-correctibility-ce0e0a3b9b4d>
- [64] James Coe and Mustafa Atay. 2021. Evaluating impact of race in facial recognition across machine learning and deep learning algorithms. *Computers* 10, 9 (2021), 113.
- [65] The New York Times Company. 2023. The New York Times Company v. OpenAI. https://nytc-co-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf Case e 1:23-cv-11195.
- [66] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070* (2018).
- [67] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- [68] Laurie Cumbo, Alicka Ampry-Samuel, Helen Rosenthal, Robert Cornegy, Ben Kallos, Adrienne Adams, Farah Louis, Margaret Chin, Fernando Cabrera, Deborah Rose, Vanessa Gibson, Justin Brannan, Carlina Rivera, Mark Levine, Diana Ayala, I. Daneek Miller, Stephen Levin, and Inez Barron. 2021. Local Law 144 of 2021. <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7cText%7c&Search=>
- [69] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. (2023). arXiv:2309.08600 [cs.LG]
- [70] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).
- [71] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. 2023. AI capabilities can be significantly improved without expensive retraining. (Dec. 2023). <https://arxiv.org/abs/2312.07413v1>
- [72] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating Data Contamination in Modern Benchmarks for Large Language Models. *arXiv preprint arXiv:2311.09783* (2023).
- [73] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *arXiv preprint arXiv:2307.08715* (2023).
- [74] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548* (2022).
- [75] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruk-sachattun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.
- [76] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2019. Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. *arXiv preprint arXiv:1911.09005* (2019).
- [77] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. 2017. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493* (2017).
- [78] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM (CACM)* (2023).
- [79] Rudresh Dwivedi, Devam Dave, Het Naik, Smriti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.
- [80] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751* (2017).
- [81] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.
- [82] Lauren B. Edelman. 1992. Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law. *Amer. J. Sociology* 97, 6 (May 1992), 1531–1576. <https://doi.org/10.1086/229939> Publisher: The University of Chicago Press.
- [83] Lauren B. Edelman. 2016. *Working Law: Courts, Corporations, and Symbolic Civil Rights*. University of Chicago Press, Chicago, IL. <https://press.uchicago.edu/ucp/books/book/chicago/W/bo24550454.html>
- [84] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics* 9 (2021), 160–175.
- [85] European Commission. 2021. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. *Eur Comm* 106 (2021), 1–108.

- [86] European Union. 2016. General Data Protection Regulation. <https://gdpr-info.eu/>
- [87] European Union. 2021. Artificial Intelligence Act. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>
- [88] European Union. 2022. Digital Markets Act. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925>
- [89] EY. 2019. EY Global Code of Conduct. Online. Retrieved from: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/generic/EY_Code_of_Conduct.pdf.
- [90] Joseph Farrell and Matthew Rabin. 1996. Cheap Talk. *Journal of Economic Perspectives* 10, 3 (Sept. 1996), 103–118. <https://doi.org/10.1257/jep.10.3.103>
- [91] Michael Feffer, Anusha Sinha, Zachary C. Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? <http://arxiv.org/abs/2401.15897> arXiv:2401.15897 [cs].
- [92] Jaden Fiotto-Kaufmann, Arnab Sen-Sharma, Caden Juang, David Bau, Eric Todd, Francesca Lucchetti, and Will Brockman. 2023. nnsight. <https://nnsight.net/>
- [93] Ross D Fuerman. 2009. Bernard Madoff and the solo auditor red flag. *Journal of Forensic & Investigative Accounting* 1, 1 (2009), 1–38.
- [94] G7. 2023. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems>
- [95] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2023. Interpreting CLIP’s Image Representation via Text-Based Decomposition. *arXiv preprint arXiv:2310.05916* (2023).
- [96] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
- [97] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
- [98] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2 (2020), 665 – 673. <https://api.semanticscholar.org/CorpusID:215786368>
- [99] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767* (2023).
- [100] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913* (2020).
- [101] Amirata Ghorbani and James Zou. 2020. Neuron Shapley: Discovering the Responsible Neurons. (2020). arXiv:2002.09815 [stat.ML]
- [102] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* 4, 6 (2022), e406–e414.
- [103] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. (Feb. 2019). <http://arxiv.org/abs/1806.00069> arXiv:1806.00069 [cs, stat].
- [104] Beng Wee Goh and Dan Li. 2013. The Disciplining Effect of the Internal Control Provisions of the Sarbanes–Oxley Act on the Governance Structures of Firms. *The International Journal of Accounting* 48, 2 (June 2013), 248–278. <https://doi.org/10.1016/j.intacc.2013.04.004>
- [105] Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493* (2023).
- [106] Arieh Goldman and Ben Zion Barlev. 1974. The Auditor-Firm Conflict of Interests: Its Implications for Independence. *The Accounting Review* 49, 4 (1974), 707–718. <https://www.jstor.org/stable/245049> Publisher: American Accounting Association.
- [107] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [108] Google. 2021. Consultation on the EU AI Act Proposal. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en
- [109] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. 2022. Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 70–88.
- [110] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. 2017. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123* (2017).
- [111] Jarek Gryz and Marcin Rojszczak. 2021. Black box algorithms and the rights of individuals: No easy solution to the “explainability” problem. *Internet Policy Review* 10, 2 (2021), 1–24.
- [112] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733* (2021).
- [113] Wes Gurnee and Max Tegmark. 2023. Language Models Represent Space and Time. (2023). arXiv:2310.02207 [cs.LG]
- [114] Philipp Hacker. 2023. The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review* 51 (Nov. 2023), 105871. <https://doi.org/10.1016/j.clsr.2023.105871>
- [115] Philipp Hacker, Johann Cordes, and Janina Rochon. 2023. Regulating Gatekeeper AI and Data: Transparency, Access, and Fairness under the DMA, the GDPR, and beyond. (Aug. 2023). <http://arxiv.org/abs/2212.04997> arXiv:2212.04997 [cs].
- [116] Philipp Hacker and Jan-Hendrik Passoth. 2022. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (Eds.). Springer International Publishing, Cham, 343–373. https://doi.org/10.1007/978-3-031-04083-2_17
- [117] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 2022. Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine* 17, 1 (Feb. 2022), 72–85. <https://doi.org/10.1109/MCI.2021.3129960> Conference Name: IEEE Computational Intelligence Magazine.
- [118] Julian Hazell. 2023. Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972* (2023).
- [119] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).
- [120] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. 2023. Foundation models and fair use. *arXiv preprint arXiv:2303.15715* (2023).
- [121] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
- [122] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural Adversarial Examples. (2021). arXiv:1907.07174 [cs.LG]
- [123] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. 2021. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.
- [124] David Hess. 2019. The Transparency Trap: Non-Financial Disclosure and the Responsibility of Business to Respect Human Rights. *American Business Law Journal* 56, 1 (2019), 5–53. <https://doi.org/10.1111/ablj.12134> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ablj.12134>
- [125] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. (2023). arXiv:2311.05232 [cs.CL]
- [126] Evan Hubinger. 2020. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532* (2020).
- [127] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024).
- [128] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*. PMLR, 2137–2146.
- [129] International Atomic Energy Agency. 2016. A Day in the Life of a Safeguards Inspector. <https://www.iaea.org/newscenter/news/a-day-in-the-life-of-a-safeguards-inspector> Accessed: 2024-04-15.
- [130] International Atomic Energy Agency. 2023. IAEA Safeguards Overview: Comprehensive Safeguards Agreements and Additional Protocols. <https://www.iaea.org/publications/factsheets/iaea-safeguards-overview>
- [131] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160* (2023).
- [132] Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786* (2023).
- [133] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
- [134] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437* (2019).

- [135] W Jeffrey Johnston and Stefano Fusi. 2023. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications* 14, 1 (2023), 1040.
- [136] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically Auditing Large Language Models via Discrete Optimization. *arXiv preprint arXiv:2303.04381* (2023).
- [137] Joemon M Jose et al. 2021. On fairness and interpretability. *arXiv preprint arXiv:2106.13271* (2021).
- [138] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4, 9 (Sept. 2023), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [139] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. *arXiv preprint arXiv:2310.13771* (2023).
- [140] Rabimba Karanjai. 2022. Targeted phishing campaigns using large scale language models. *arXiv preprint arXiv:2301.00665* (2022).
- [141] Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora, Adam Dziedzic, Franziska Boenisch, Tom Brown, Jacob Steinhardt, and Dan Hendrycks. 2023. Testing Robustness Against Unforeseen Adversaries. (2023). arXiv:1908.08016 [cs.LG]
- [142] Emre Kazim, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. 2021. Systematizing Audit in Algorithmic Recruitment. *Journal of Intelligence* 9, 3 (Sept. 2021), 46. <https://doi.org/10.3390/jintelligence9030046> Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- [143] Ashraf Khalil, Soha Glal Ahmed, Asad Masood Khattak, and Nabeel Al-Qirim. 2020. Investigating Bias in Facial Analysis Systems: A Systematic Review. *IEEE Access* 8 (2020), 130751–130761. <https://doi.org/10.1109/ACCESS.2020.3006051> Conference Name: IEEE Access.
- [144] Mohd Ehmer Khan and Farmeena Khan. 2012. A comparative study of white box, black box and grey box testing techniques. *International Journal of Advanced Computer Science and Applications* 3, 6 (2012).
- [145] Heidy Khlaaf. 2023. How AI Can Be Regulated Like Nuclear Energy. *TIME* (Oct. 2023). <https://time.com/6327635/ai-needs-to-be-regulated-like-nuclear-weapons/>
- [146] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html> ISSN: 2640-3498.
- [147] 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. 2023. Evaluating Language-Model Agents on Realistic Autonomous Tasks. <https://evals.alignment.org/language-model-pilot-report>. (July 2023).
- [148] Shunsuke Kitada and Hitoshi Iyatomi. 2023. Making attention mechanisms more robust and interpretable with virtual adversarial training. *Applied Intelligence* 53, 12 (2023), 15802–15817.
- [149] Leonie Koessler and Jonas Schuett. 2023. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. (July 2023). <https://arxiv.org/abs/2307.08823v1>
- [150] Noam Kolt. 2023. Algorithmic black swans. *Washington University Law Review* 101 (2023).
- [151] Adriano Koshiyama, Emre Kazim, and Philip Treleaven. 2022. Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *Computer* 55, 4 (2022), 40–50.
- [152] Kimberly D. Krawiec. 2003. Cosmetic Compliance and the Failure of Negotiated Governance. *SSRN Electronic Journal* (2003). <https://doi.org/10.2139/ssrn.448221>
- [153] Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Understanding the Effects of Iterative Prompting on Truthfulness. *CoRR* abs/2402.06625 (2024). <https://doi.org/10.48550/ARXIV.2402.06625> arXiv:2402.06625
- [154] Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Prukshatkatun, and Kai-Wei Chang. 2022. Measuring Fairness of Text Classifiers via Prediction Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5830–5842.
- [155] Maya Krishnan. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* 33, 3 (2020), 487–502.
- [156] Yilun Kuang and Yash Bharti. [n. d.]. Scale-invariant-Fine-Tuning (SiFT) for Improved Generalization in Classification. ([n. d.]).
- [157] Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2251–2277.
- [158] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N Balasubramanian. 2019. Harnessing the vulnerability of latent layers in adversarially trained models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2779–2785.
- [159] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. Entangled Preferences: The History and Risks of Reinforcement Learning and Human Feedback. (2023). arXiv:2310.13595 [cs.CY]
- [160] Richard N Landers and Tara S Behrend. 2023. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist* 78, 1 (2023), 36.
- [161] Jose Antonio Lanz. 2023. Stable Diffusion XL v0.9 Leaks Early, Generating Raves From Users. <https://decrypt.co/147612/stable-diffusion-xl-v0-9-leaks-early-generating-raves-from-users>
- [162] Raz Lapid, Ron Langberg, and Moshe Sipper. 2023. Open Sesame! Universal Black Box Jailbreaking of Large Language Models. *arXiv preprint arXiv:2309.01446* (2023).
- [163] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? , 138–138 pages.
- [164] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. *arXiv preprint arXiv:2401.01967* (2024).
- [165] Sharkey Lee, Ghuidhir Clíodhna Ní, Dan Braun, Scheurer Jérémy, Mikita Balesni, Bushnaq Lucius, Stix Charlotte, and Marius Hobbhahn. 2023. A causal framework for AI Regulation and Auditing. (2023).
- [166] Mark A. Lemley, Peter Henderson, and Tatsunori Hashimoto. 2023. Where's the Liability in Harmful AI Speech? *SSRN Electronic Journal* (2023). <https://doi.org/10.2139/ssrn.4531029>
- [167] Clive Lennox. 2000. Do companies successfully engage in opinion-shopping? Evidence from the UK. *Journal of Accounting and Economics* 29, 3 (June 2000), 321–337. [https://doi.org/10.1016/S0165-4101\(00\)00025-2](https://doi.org/10.1016/S0165-4101(00)00025-2)
- [168] Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. (2023). arXiv:2310.20624 [cs.LG]
- [169] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023. Multi-step Jailbreaking Privacy Attacks on ChatGPT. *arXiv preprint arXiv:2304.05197* (2023).
- [170] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271* (2018).
- [171] Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8410–8418.
- [172] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
- [173] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.
- [174] Cheryl Linthicum, Austin L Reitenga, and Juan Manuel Sanchez. 2010. Social responsibility and corporate reputation: The case of the Arthur Andersen Enron audit failure. *Journal of Accounting and Public Policy* 29, 2 (2010), 160–176.
- [175] Aiwei Liu, Honghai Yu, Xuming Hu, Shuang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. 2022. Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution. *ArXiv* abs/2210.17004 (2022). <https://api.semanticscholar.org/CorpusID:253236900>
- [176] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994* (2020).
- [177] Xiaoxuan Liu, Ben Glocker, Melissa M. McCadden, Marzyeh Ghassemi, Alastair K. Denniston, and Lauren Oakden-Rayner. 2022. The medical algorithmic audit. *The Lancet Digital Health* 4, 5 (May 2022), e384–e397. [https://doi.org/10.1016/S2589-7500\(22\)00003-6](https://doi.org/10.1016/S2589-7500(22)00003-6) Publisher: Elsevier.
- [178] Xingbin Liu, Huafeng Kuang, Hong Liu, Xianming Lin, Yongjian Wu, and Rongrong Ji. 2023. Latent Feature Relation Consistency for Adversarial Robustness. *arXiv preprint arXiv:2303.16697* (2023).
- [179] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).
- [180] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860* (2023).
- [181] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. 2023. AI Regulation Is (not) All You Need. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023). <https://api.semanticscholar.org/CorpusID:259139804>
- [182] Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv preprint arXiv:2105.02732* (2021).
- [183] Vidur Mahajan, Vasantha Kumar Venugopal, Murali Murugavel, and Harsh Mahajan. 2020. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms—How We Do It. *Academic Radiology* 27, 1 (Jan. 2020), 132–135. <https://doi.org/10.1016/j.acra.2019.09.009>

- [184] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. *arXiv:2403.19647* [cs.LG].
- [185] Samuel Marks and Max Tegmark. 2023. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. (2023). *arXiv:2310.06824* [cs.AI].
- [186] Christopher Marquis, Michael W. Toffel, and Yanhua Zhou. 2016. Scrutiny, Norms, and Selective Disclosure: A Global Study of Greenwashing. *Organization Science* 27, 2 (March 2016), 483–504. <https://doi.org/10.1287/orsc.2015.1039> Publisher: INFORMS.
- [187] Miljan Martić, Jan Leike, Andrew Trask, Matteo Hessel, Shane Legg, and Pushmeet Kohli. 2018. Scaling shared model governance via model splitting. *arXiv preprint arXiv:1812.05979* (2018).
- [188] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.
- [189] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems* 35 (2022), 17359–17372.
- [190] Jacob Metcalfe, Emanuel Moss, Ranjit Singh, Emnet Tafese, and Elizabeth Anne Watkins. 2022. A relationship and not a thing: A relational approach to algorithmic accountability and assessment documentation. *arXiv preprint arXiv:2203.01455* (2022).
- [191] Jacob Metcalfe, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.
- [192] METR. 2023. METR. <https://evals.alignment.org/>
- [193] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [194] Andrea Miotti and Akash Wasil. 2023. Taking control: Policies to address extinction risks from advanced AI. *arXiv preprint arXiv:2310.20563* (2023).
- [195] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [196] Jakob Mökander. 2023. Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society* 2, 3 (2023), 49.
- [197] Don A. Moore, Philip E. Tetlock, Lloyd Tanlu, and Max H. Bazerman. 2006. Conflicts of Interest and the Case of Auditor Independence: Moral Seduction and Strategic Issue Cycling. *The Academy of Management Review* 31, 1 (2006), 10–29. <https://www.jstor.org/stable/20159182> Publisher: Academy of Management.
- [198] Christopher A Mouton, Caleb Lucas, and Ella Guest. 2023. The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach. (2023).
- [199] Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Advances in Neural Information Processing Systems* 33 (2020), 17153–17163.
- [200] Jakob Mökander and Luciano Floridi. 2021. Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines* 31, 2 (June 2021), 323–327. <https://doi.org/10.1007/s11023-021-09557-8>
- [201] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics* (may 2023). <https://doi.org/10.1007/s43681-023-00289-2>
- [202] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics* (May 2023). <https://doi.org/10.1007/s43681-023-00289-2> *arXiv:2302.08500* [cs].
- [203] Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. 2023. Testing Language Model Agents Safely in the Wild. (Dec. 2023). <https://doi.org/10.48550/arXiv.2311.10538> *arXiv:2311.10538* [cs].
- [204] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. (2023). *arXiv:2301.05217* [cs.LG].
- [205] Arvind Narayanan and Sayash Kapoor. 2023. Evaluating LLMs is a minefield. https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/#/8
- [206] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. (Nov. 2023). <https://doi.org/10.48550/arXiv.2311.17035> *arXiv:2311.17035* [cs].
- [207] National Institute for Standards and Technology. 2023. Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11). <https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the>
- [208] National New Generation Artificial Intelligence Governance Expert Committee. 2019. Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'. <https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/>
- [209] National New Generation Artificial Intelligence Governance Specialist Committee. 2021. "Ethical Norms for New Generation Artificial Intelligence" Released. <https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/>
- [210] Sella Nevo, Dan Lahav, Ajay Karapur, Jeff Alstott, and Jason Matheny. 2023. *Securing Artificial Intelligence Model Weights: Interim Report*. Technical Report. RAND Corporation. https://www.rand.org/pubs/working_papers/WRA2849-1.html
- [211] Kwan Yee Ng, Jason Zhou, Ben Murphy, Rogier Creemers, and Hunter Dorwart. 2023. Translation: Artificial Intelligence Law, Model Law v. 1.0 (Expert Suggestion Draft) – Aug. 2023. (Aug. 2023). <https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/>
- [212] Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626* (2022).
- [213] Aaron L Nielson. 2018. Sticky Regulations. *U. Chi. L. Rev* 85 (2018), 85.
- [214] OECD. 2019. Recommendation of the Council on Artificial Intelligence. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [215] Electronic Code of Federal Regulations. 2023. Regulation M. Code of Federal Regulations. <https://www.ecfr.gov/current/title-17/chapter-II/part-242/subject-group-ECFR3dd95cf4d3f6730> 17 CFR Part 242.
- [216] Office of Science and Technology Policy. 2022. Notice and Explanation. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/notice-and-explanation/>
- [217] Office of the President of the United States. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. <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/>
- [218] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *arXiv preprint arXiv:2402.17861* (2024).
- [219] Christine Oliver. 1991. Strategic responses to institutional processes. *Academy of Management Review* 16, 1 (Jan. 1991), 145–179. <https://doi.org/10.5465/amr.1991.4279002> Publisher: Academy of Management.
- [220] A.J. Oneal. 2023. Chat GPT "DAN" (and other "Jailbreaks"). <https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516>.
- [221] OpenAI. 2023. GPT-3.5 Turbo fine-tuning and API updates. <https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates>
- [222] OpenAI. 2023. GPT-4 Technical Report. (2023). *arXiv:2303.08774* [cs.CL].
- [223] OpenAI. 2023. OpenAI Preparedness Challenge. <https://openai.com/form/preparedness-challenge>
- [224] OpenAI. 2023. OpenAI Red Teaming Network. <https://openai.com/blog/red-teaming-network>
- [225] Openminded. 2023. How to audit an AI model owned by someone else (part 1). *OpenMinded Blog* (June 2023). <https://blog.openminded.org/ai-audit-part-1/>
- [226] Genki Osada, Budrul Ahsan, Revoti Prasad Bora, and Takashi Nishide. 2022. Latent Space Virtual Adversarial Training for Supervised and Semi-Supervised Learning. *IEICE TRANSACTIONS on Information and Systems* 105, 3 (2022), 667–678.
- [227] Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11130–11138.
- [228] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. 2016. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768* (2016).
- [229] Geon Yeong Park and Sang Wan Lee. 2021. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7758–7767.
- [230] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. (Aug. 2023). <https://doi.org/10.48550/arXiv.2304.03442> *arXiv:2304.03442* [cs].
- [231] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. (2023). *arXiv:2308.14752* [cs.CY].
- [232] PCAOB. 2002. Sarbanes-Oxley Act of 2002. https://pcaobus.org/About/History/Documents/PDFs/Sarbanes_Oxley_Act_of_2002.pdf Public Law 107-204, 116 Stat. 745.
- [233] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).
- [234] Ethan Perez, Sam Ringer, Kamille Lukošiuaitė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251* (2022).

- [235] Personal Data Protection Commission Singapore. 2020. Model Artificial Intelligence Governance Framework, Second Edition. <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf>
- [236] P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. *Four principles of explainable artificial intelligence*. Technical Report NIST IR 8312. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. NIST IR 8312 pages. <https://doi.org/10.6028/NIST.IR.8312>
- [237] P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. 2021. *Four Principles of Explainable Artificial Intelligence*. Interagency or Internal Report 8312. National Institute for Standards and Technology.
- [238] Thomas Ploug and Soren Holm. 2021. Right to Contest AI Diagnostics: Defining Transparency and Explainability Requirements from a Patient’s Perspective. In *Artificial Intelligence in Medicine*. Springer, 1–12.
- [239] Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research. *arXiv preprint arXiv:2304.12397* (2023).
- [240] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281* (2022).
- [241] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual Adversarial Examples Jailbreak Large Language Models. *arXiv preprint arXiv:2306.13213* (2023).
- [242] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv preprint arXiv:2310.03693* (2023).
- [243] Yaguan Qian, Qiqi Shao, Tengting Yao, Bin Wang, Shouling Ji, Shaoning Zeng, Zhaoquan Gu, and Wassim Swaileh. 2021. Towards Speeding up Adversarial Training in Latent Spaces. *arXiv preprint arXiv:2102.00662* (2021).
- [244] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873* (2023).
- [245] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20)*. Association for Computing Machinery, New York, NY, USA, 469–481. <https://doi.org/10.1145/3351095.3372828>
- [246] Manish Raghavan and Pauline Kim. 2023. Limitations of the “Four-Fifths Rule” and Statistical Parity Tests for Measuring Fairness. [https://openreview.net/forum?id=M2aNjwX4Ec&referrer=%5Bthe%20profile%20of%20Manish%20Raghavan%5D\(%2Fprofile%3Fid%3D-Manish_Raghavan1\)](https://openreview.net/forum?id=M2aNjwX4Ec&referrer=%5Bthe%20profile%20of%20Manish%20Raghavan%5D(%2Fprofile%3Fid%3D-Manish_Raghavan1))
- [247] Inioluwa Deborah Raji. 2022. The Anatomy of AI Audits: Form, Process, and Consequences. (2022).
- [248] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.
- [249] Inioluwa Deborah Raji and Joy Buolamwini. 2022. Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Commun. ACM* 66, 1 (2022), 101–108.
- [250] Inioluwa Deborah Raji, Timmit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AI/ES ’20)*. Association for Computing Machinery, New York, NY, USA, 145–151. <https://doi.org/10.1145/3375627.3375820>
- [251] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timmit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 33–44. <https://doi.org/10.1145/3351095.3372873>
- [252] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.
- [253] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610* (2022).
- [254] Javier Rando and Florian Tramèr. 2023. Universal Jailbreak Backdoors from Poisoned Human Feedback. (2023). arXiv:2311.14455 [cs.AI]
- [255] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. (2023). arXiv:2305.14965 [cs.CL]
- [256] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 464–483.
- [257] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719* (2020).
- [258] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1085–1097.
- [259] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).
- [260] Ronald E. Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference (WWW ’18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 955–965. <https://doi.org/10.1145/3178876.3186143>
- [261] Daniel Rodriguez Maffioli. 2023. Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency. Available at SSRN 4579322 (2023).
- [262] Emma Roth. 2023. The New York Times is suing OpenAI and Microsoft for copyright infringement. *The Verge* (Dec. 2023). <https://www.theverge.com/2023/12/27/24016212/new-york-times-openai-microsoft-lawsuit-copyright-infringement>
- [263] Tom Roth, Yansong Gao, Alsharif Abuadba, Surya Nepal, and Wei Liu. 2021. Token-Modification Adversarial Attacks for Natural Language Processing: A Survey. *ArXiv abs/2103.00676* (2021). <https://api.semanticscholar.org/CorpusID:232075640>
- [264] Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *Stat* 1050 (2018), 26.
- [265] Teerapong Sae-Lim and Suronapee Phoomvuthisarn. 2022. Weighted Token-Level Virtual Adversarial Training in Text Classification. In *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*. IEEE, 117–123.
- [266] Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952* (2023).
- [267] Swami Sankaranarayanan, Arpit Jain, Rama Chellappa, and Ser Nam Lim. 2018. Regularizing deep networks using efficient layerwise adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [268] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548* (2023).
- [269] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.
- [270] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? (2023). arXiv:2304.15004 [cs.AI]
- [271] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2023. Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. *arXiv preprint arXiv:2311.07590* (2023).
- [272] Jonas Schuett. 2022. Three lines of defense against risks from AI. *arXiv preprint arXiv:2212.08364* (2022).
- [273] Jonas Schuett. 2023. AGI labs need an internal audit function. (May 2023). <https://arxiv.org/abs/2305.17038v1>
- [274] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153* (2023).
- [275] Leo Schwinn, David Dobre, Stephan Günemann, and Gauthier Gidel. 2023. Adversarial Attacks and Defenses in Large Language Models: Old and New Threats. (2023). arXiv:2310.19737 [cs.AI]
- [276] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. (2023).
- [277] Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. 2023. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. (2023). arXiv:2311.03348 [cs.CL]
- [278] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *Comput. Surveys* (2023).
- [279] Lee Sharkey, Clíodhna Ní Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn. 2024. A Causal Framework for AI Regulation and Auditing. (2024).

- [280] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. (2023). arXiv:2310.13548 [cs.CL]
- [281] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. *arXiv preprint arXiv:2310.10844* (2023).
- [282] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *arXiv preprint arXiv:2308.03825* (2023).
- [283] Toby Shevlane. 2022. Structured access: an emerging paradigm for safe AI deployment. (2022). arXiv:2201.05159 [cs.AI]
- [284] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324* (2023).
- [285] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789* (2023).
- [286] Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2022. Toward Human Readable Prompt Tuning: Kubrick's The Shining is a good movie, and a good prompt too? *arXiv preprint arXiv:2212.10539* (2022).
- [287] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).
- [288] Michal Shur-Ofry. 2023. Multiplicity as an AI Governance Principle. Available at SSRN 4444354 (2023).
- [289] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138* (2022).
- [290] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. <https://doi.org/10.1145/3375627.3375830>
- [291] Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2023. Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey. *arXiv preprint arXiv:2310.01424* (2023).
- [292] Emily H Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? *arXiv preprint arXiv:2306.03809* (2023).
- [293] Irene Solaiman. 2023. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 111–122.
- [294] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949* (2023).
- [295] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2020. Universal adversarial attacks with natural triggers for text classification. *arXiv preprint arXiv:2005.00174* (2020).
- [296] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofer Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* (2024).
- [297] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).
- [298] Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks. (2023). arXiv:2311.09730 [cs.CL]
- [299] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhao Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kaikhura, Caiming Xiong, Chao Zhang, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyi Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:2401.05561 [cs.CL]
- [300] Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2023. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. *arXiv preprint arXiv:2305.04400* (2023).
- [301] Wesley Tann, Yuancheng Liu, Jun Heng Sim, Choon Meng Seah, and Ee-Chien Chang. 2023. Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions. *arXiv preprint arXiv:2308.10443* (2023).
- [302] Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2023. Auditing and Mitigating Cultural Bias in LLMs. (2023). arXiv:2311.14096 [cs.CL]
- [303] Max Tegmark and Steve Omohundro. 2023. Provably safe systems: the only path to controllable AGI. (Sept. 2023). <https://doi.org/10.48550/arXiv.2309.01933> [cs].
- [304] David Thiel. 2023. Identifying and Eliminating CSAM in Generative ML Training Data and Models. (2023).
- [305] David Thiel, Melissa Stroebel, and Rebecca Portnoff. 2023. Generative ML and CSAM: Implications and Mitigations. (2023).
- [306] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023). arXiv:2307.09288 [cs.CL]
- [307] Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al. 2023. International governance of civilian AI: A jurisdictional certification approach. *arXiv preprint arXiv:2308.15514* (2023).
- [308] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2023. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *arXiv preprint arXiv:2310.10012* (2023).
- [309] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248* (2023).
- [310] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. (2023). arXiv:2305.04388 [cs.CL]
- [311] UK Department for Science, Innovation & Technology. 2023. *A pro-innovation approach to AI regulation*. Technical Report. <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>
- [312] United Nations. 2022. Principles for the ethical use of artificial intelligence in the United Nations system. https://unsceb.org/sites/default/files/2023-03/CEB_2022_2_Add.1%20%28AI%20ethics%20principles%29.pdf
- [313] United States National Science Foundation. 2023. National Deep Inference Facility for Very Large Language Models (NDIF). (2023).
- [314] U.S. Department of Commerce and National Institute of Standards and Technology. 2023. AI Risk Management Framework: AI RMF (1.0). <https://doi.org/10.6028/NIST.AI.100-1>
- [315] H. E. van den Brom. 2022. On-site Inspection and Legal Certainty. *SSRN Electronic Journal* (2022). <https://api.semanticscholar.org/CorpusID:249326468>
- [316] Stephen Wagner and Lee Dittmar. 2006. The unexpected benefits of Sarbanes-Oxley. *Harvard Business Review* 84, 4 (April 2006), 133–140; 150.
- [317] Ari Ezra Waldman. 2019. Privacy Law's False Promise. *SSRN Electronic Journal* (2019). <https://doi.org/10.2139/ssrn.3339372>
- [318] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125* (2019).
- [319] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning Language Models During Instruction Tuning. (2023). arXiv:2305.00944 [cs.CL]
- [320] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
- [321] Jiong Xiao Wang, Junlin Wu, Muhan Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. 2023. On the Exploitability of Reinforcement Learning with Human Feedback for Large Language Models. (2023). arXiv:2311.09641 [cs.AI]
- [322] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023. Knowledge Editing for Large Language Models: A Survey. (2023). arXiv:2310.16218 [cs.CL]

- [323] Tony T. Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell. 2023. Adversarial Policies Beat Superhuman Go AIs. (2023). arXiv:2211.00241 [cs.LG]
- [324] Elizabeth Anne Watkins, Emanuel Moss, Jacob Metcalf, Ranjit Singh, and Madeleine Clare Elish. 2021. Governing algorithmic systems with impact assessments: Six observations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 1010–1022.
- [325] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483* (2023).
- [326] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [327] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
- [328] 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. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. (Oct. 2023). <http://arxiv.org/abs/2310.11986> arXiv:2310.11986 [cs].
- [329] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668* (2023).
- [330] Evan Westra. 2021. Virtue Signaling and Moral Progress. *Philosophy & Public Affairs* 49, 2 (2021), 156–178. <https://doi.org/10.1111/papa.12187> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/papa.12187>.
- [331] Lawrence J. White. 2010. Markets: The Credit Rating Agencies. *Journal of Economic Perspectives* 24, 2 (June 2010), 211–226. <https://doi.org/10.1257/jep.24.2.211>
- [332] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 1–18.
- [333] Daricia Wilkinson, Kate Crawford, Hanna Wallach, Deborah Raji, Bogdana Rakova, Ranjit Singh, Angelika Strohmayer, and Ethan Zuckerman. 2023. Accountability in Algorithmic Systems: From Principles to Practice. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.
- [334] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Chao Shen, and Hongyuan Zha. 2022. BackdoorBench: A Comprehensive Benchmark of Backdoor Learning. *arXiv preprint arXiv:2206.12654* (2022).
- [335] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and Editing Privacy Neurons in Pretrained Language Models. (2023). arXiv:2310.20138 [cs.CR]
- [336] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. (2023). arXiv:2310.02949 [cs.CL]
- [337] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. 2023. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *arXiv preprint arXiv:2312.02003* (2023).
- [338] Rui-Jie Yew and Dylan Hadfield-Menell. 2022. A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 823–830.
- [339] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-Resource Languages Jailbreak GPT-4. (2023). arXiv:2310.02446 [cs.CL]
- [340] Jiahao Yu, Kingwei Lin, and Xinyu Xing. 2023. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv preprint arXiv:2309.10253* (2023).
- [341] Mert Yuksekgonul, Maggie Wang, and James Zou. 2022. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480* (2022).
- [342] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2023. Removing RLHF Protections in GPT-4 via Fine-Tuning. (2023). arXiv:2311.05553 [cs.CL]
- [343] Chanyuan Abigail Zhang, Soohyun Cho, and Miklos Vasarhelyi. 2022. Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems* 46 (2022), 100572.
- [344] Milin Zhang, Mohammad Abdi, and Francesco Restuccia. 2023. Adversarial Machine Learning in Latent Representations of Neural Networks. *arXiv preprint arXiv:2309.17401* (2023).
- [345] W. Zhang, Quan Z Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. 2019. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. *arXiv: Computation and Language* (2019). <https://api.semanticscholar.org/CorpusID:260428188>
- [346] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.
- [347] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* (2023).
- [348] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks. (2023). arXiv:2306.17844 [cs.LG]
- [349] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. WebArena: A Realistic Web Environment for Building Autonomous Agents. (Oct. 2023). <https://doi.org/10.48550/arXiv.2307.13854> arXiv:2307.13854 [cs].
- [350] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. 2018. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 452–467.
- [351] Xiaowei Zhou, Ivor W Tsang, and Jie Yin. 2019. Latent adversarial defence with boundary-guided generation. *arXiv preprint arXiv:1907.07001* (2019).
- [352] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. FreeLb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764* (2019).
- [353] Daniel M. Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmid-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. 2022. Adversarial Training for High-Stakes Reliability. (2022). arXiv:2205.01663 [cs.LG]
- [354] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).
- [355] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. (July 2023). <https://doi.org/10.48550/arXiv.2307.15043> arXiv:2307.15043 [cs].

A MOTIVATIONS FOR EXTERNAL AUDITS

Audits involve formally evaluating systems to assess risks, compliance with standards and regulations, and other desiderata of interest to stakeholders. High-quality audits from independent, external auditors have been motivated in multiple ways [279]:

- **Identifying problems:** The most direct purpose of audits is to identify risks from unsound systems or practices.
- **Incentivizing responsible development:** When individual components of the development process are insufficiently documented, information necessary to contextually assess risks is lost [145]. Audits can assess the sufficiency of internal controls, risk assessment, and documentation [33, 149, 251, 273]. Greater accountability for internal practices incentivizes auditees to spend more effort on risk mitigation and documentation [316], especially when facing penalties or public scrutiny [104, 338].
- **Increasing transparency:** Publicly shared information from audits can help regulators and the scientific community develop a better understanding of system behaviors and limitations.
- **Enabling fixes to technical problems:** When problems are found during an audit, developers can then work to address them [248]. External audits can also identify risk factors that might merit further guardrails on deployment, closer monitoring of deployed systems, or follow-up studies of user impacts.
- **Balancing transparency and security:** Keeping systems entirely secret is maximally secure but prevents external scrutiny. Open-sourcing them allows for maximal scrutiny but can proliferate proprietary or misusable systems [276]. Audits offer a middle ground that allows for some transparency and independent risk assessment with high security.
- **Providing greater credibility to responsible developers:** Passing audits increases trust in developers and their systems. Hence, the public can better calibrate their trust in developers and systems.

B TECHNICAL ASSISTANCE AS A FORM OF OUTSIDE-THE-BOX ACCESS

In Section 5, we discuss how outside-the-box access to information can help auditors conduct audits more effectively. However, for similar reasons, access to technical assistance can also be useful. For example, one resource that auditors will often need, especially for large language models, is computing infrastructure [11, 218]. Further, additional technical assistance from the developers' engineers may also help because they have unique practical knowledge of working effectively with their models. This may include assistance with fine-tuning, developing realistic test cases (e.g., [349]), or integrating models with external tools that enhance capabilities to resemble real-world usage [71, 165, 203, 230]. Past experience with AI audits has highlighted the value of technical assistance from developers.

After seeing the final audit report, we realized that we could have helped [METR, (formerly ARC Evals)] be more successful in identifying concerning behavior if we had known more details about their (clever and

well-designed) audit approach. This is because getting models to perform near the limits of their capabilities is a fundamentally difficult research endeavor. Prompt engineering and fine-tuning language models are active research areas, with most expertise residing within AI companies. With more collaboration, we could have leveraged our deep technical knowledge of our models to help [METR] execute the evaluation more effectively.
 –Anthropic on their audit by METR (formerly ARC Evals) [13]

Allowing developers to arbitrarily influence audits undermines their independence, so incorporating requirements for developers to provide technical assistance into legal auditing frameworks may be difficult and is beyond the scope of this paper. However, auditors may find it helpful if specific requests for technical assistance are answered in good faith by auditees.

C INNOVATION ON AUDITING TOOLS

White-box tools for studying AI systems have long been a topic of technical interest, but research on methods often struggles to keep up with the scale and capabilities of leading AI systems [20]. There are gaps between the capabilities of white-box evaluation tools and what auditors may need from them. More progress on both foundational research and practical tools will be useful for auditors, especially for state-of-the-art large language models because of their unique versatility and complexity.

Basic research: Current methods have provided useful insights. However, developing a detailed mechanistic understanding is not yet possible in state-of-the-art models. More progress in the basic science of neural networks and efforts to study their inner workings will help further research on evaluation techniques. This will require progress on both developing more intrinsically understandable systems and techniques to interpret trained ones [45, 256].

Practical tools: The goal of research on evaluation techniques is to produce methods that can be effectively used off-the-shelf by auditors. In the adversarial attack literature, benchmarks have largely focused on fooling networks with small perturbations to inputs instead of eliciting harm via more real-world features [122, 141, 263, 345]. In the interpretability literature, few benchmarks connected to practical tasks exist, with it being common to judge techniques based on researcher intuition [3, 50, 77, 155, 193, 256]. Given the increasing scale and complexity of modern AI systems, developing more effective evaluation tools poses a challenge. Fortunately, open-source and API access to advanced AI systems has enabled progress on evaluation tools. However, no technique for benchmarking evaluation tools is more directly informative than applications on real systems. Partnerships between researchers and developers can facilitate these.

Secure auditing infrastructure: As discussed in Section 6, granting auditors white-box access to systems via application programming interfaces or secure research environments can reduce the risk of leaks. However, because norms for AI audits have not yet been established, there is little infrastructure for conducting audits securely. For example, efforts like the US National Deep Inference Facility project [313] could make more resources available to auditors. Establishing better tools and protocols is another priority

[181]. At the same time, it will be key to establish norms and a regulatory framework around AI audits, as has been done in other industries with audits.

D BEYOND ACCESS: OTHER ASPECTS OF RIGOROUS AUDITS.

White- and outside-the-box access is necessary but not sufficient for rigorous audits. Many factors can undermine or degrade the quality of audits. We overview challenges here.

Poorly-resourced audits: Working with state-of-the-art AI systems and effectively evaluating them requires compute and technical expertise. While developing and commercializing advanced AI systems can be lucrative, searching for problems with them might not be profitable or financially sustainable. Existing audits have largely relied on private funding (e.g., [13, 147, 192, 222]), rather than public funding or other more sustainable, reliable, and diversified sources of funding.

Limitations with technical tools: As discussed in Appendix C, there is a gap between existing technical tools for evaluations and the kind of tooling needed to reliably assess the safety and trustworthiness of advanced systems. Until this gap is closed, audits will be limited in identifying risks.

Narrowly-scoped audits: Audits may omit important evaluations. For example, early audits of GPT-4 have focused on risk-related capabilities [147, 192, 222] but did not appear to include external evaluation regarding other concerns such as robustness to adversarial attacks; potential for misinformation; demographic representation; or impacts on societal welfare, democracy, discrimination, and equality. Another way in which auditing can be narrow in scope is if it only occurs pre-deployment. A “black cloud” system with ever-changing components is even more difficult to evaluate than a black box.

Conflicts of interest: Auditors may face pressure to refrain from insisting on sufficient access or conducting sufficiently rigorous audits. Auditor conflicts of interest, including collusion with auditees [27, 331] are well-known and long-standing problems [106, 197]. They stem in part from the typical payment structure of auditors: auditors that produce more favorable evaluations – including due to receiving inadequate or incomplete information from audit targets – are often preferred over other auditors, leading companies to “opinion-shop” [167] for comparatively lax evaluations. This can trigger a race to the bottom in which audits become progressively less rigorous and less informative [11, 17]. This type of dynamic could emerge in the absence of adequate regulatory structures. For example, recent audits of state-of-the-art language models from OpenAI [222] and Anthropic [13] were conducted on a voluntary basis by the Model Evaluation and Threat Research organization (METR, formerly named ARC Evals) [192], which maintains a close relationship with both companies the details of which are not publicly disclosed.

Exclusion of under-represented viewpoints: Not all people agree on what behaviors from AI systems are harmful. As a result, audits can exclude under-represented groups if they are designed in a way that fails to take a wide range of interests into account [22, 91, 163]. Improving meaningful participation [288, 296] and

dialogue [76] among diverse stakeholders plays an indispensable role in improving fairness and representation.

Cosmetic compliance: Absent clear legal requirements, companies have an incentive to prioritize cosmetic compliance with good practices [152], a form of cheap talk [90] or virtue signaling [330] in which audit targets create a superficial (yet misleading) appearance of good faith cooperation.

Regulatory capture: While governance regimes that bolster auditing standards and procedures may appear promising, they, too, can be undermined. Studies in the field of organizational science demonstrate that companies respond strategically to interventions, employing a variety of operational, political, and legal tactics [219] including supporting biased research [1]. In its simplest form, companies may selectively disclose audit-relevant information [124, 186], enabling them to game outcomes, including in AI audits [25, 252]. Meanwhile, more sophisticated and well-resourced companies can shape the underlying audit criteria, metrics, and institutions, including by selecting which auditors have privileged access to information and which do not. Legal sociologists describe this symbiotic relationship between regulators and regulated entities as “legal endogeneity”: it is precisely the actors that law seeks to control that end up controlling the law [1, 82, 83, 317]. AI audits are especially susceptible to these dynamics because the relevant standards are currently unclear [67] and audit tools are bespoke and applied inconsistently across different developers and domains [67].