AI and Ethics (2024) 4:1085–1115
https://doi.org/10.1007/s43681-023-00289-2

ORIGINAL RESEARCH

Auditing large language models: a three-layered approach

Jakob Mökander[1,2] · Jonas Schuett[3,4] · Hannah Rose Kirk[1] · Luciano Floridi[1,5]

Received: 17 February 2023 / Accepted: 18 April 2023 / Published online: 30 May 2023

Abstract

Large language models (LLMs) represent a major advance in artificial intelligence (AI) research. However, the widespread use of LLMs is also coupled with significant ethical and social challenges. Previous research has pointed towards auditing as a promising governance mechanism to help ensure that AI systems are designed and deployed in ways that are ethical, legal, and technically robust. However, existing auditing procedures fail to address the governance challenges posed by LLMs, which display emergent capabilities and are adaptable to a wide range of downstream tasks. In this article, we address that gap by outlining a novel blueprint for how to audit LLMs. Specifically, we propose a three-layered approach, whereby governance audits (of technology providers that design and disseminate LLMs), model audits (of LLMs after pre-training but prior to their release), and application audits (of applications based on LLMs) complement and inform each other. We show how audits, when conducted in a structured and coordinated manner on all three levels, can be a feasible and effective mechanism for identifying and managing some of the ethical and social risks posed by LLMs. However, it is important to remain realistic about what auditing can reasonably be expected to achieve. Therefore, we discuss the limitations not only of our three-layered approach but also of the prospect of auditing LLMs at all. Ultimately, this article seeks to expand the methodological toolkit available to technology providers and policymakers who wish to analyse and evaluate LLMs from technical, ethical, and legal perspectives.

Keywords Artificial intelligence · Auditing · Ethics · Foundation models · Governance · Large language models · Natural language processing · Policy · Risk management

1 Introduction

Auditing is a governance mechanism that technology providers and policymakers can use to identify and mitigate risks associated with artificial intelligence (AI) systems [1-5].[1] Auditing is characterised by a systematic and independent process of obtaining and evaluating evidence regarding an entity's actions or properties and communicating the results of that evaluation to relevant stakeholders [6]. Three ideas underpin the promise of auditing as an AI governance mechanism: that procedural regularity and transparency contribute to good governance [7, 8]; that proactivity in the design of AI systems helps identify risks and prevent harm before it occurs [9, 10]; and, that the operational independence between the auditor and the auditee contributes to the objectivity and professionalism of the evaluation [11, 12].

✉ Jakob Mökander
jakob.mokander@oii.ox.ac.uk

Jonas Schuett
jonas.schuett@governance.ai

Hannah Rose Kirk
hannah.kirk@oii.ox.ac.uk

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, Oxford, UK

[2] Center for Information Technology Policy, Princeton University, Princeton, USA

[3] Centre for the Governance of AI, Oxford, UK

[4] Faculty of Law, Goethe University Frankfurt, Frankfurt am Main, Germany

[5] Department of Legal Studies, University of Bologna, Bologna, Italy

[1] The term *governance mechanism* refers to the set of activities and controls used by various parties in society to exert influence and achieve normative ends [275].

Previous work on AI auditing has focused on ensuring that specific applications meet predefined, often sector-specific, requirements. For example, researchers have developed procedures for how to audit AI systems used in recruitment [13], online search [14], image classification [15], and medical diagnostics [16, 17]. However, the capabilities of AI systems tend to become ever more general. In a recent article, Bommasani et al. [18] coined the term *foundation models* to describe models that can be adapted to a wide range of downstream tasks. While foundation models are not necessarily new from a technical perspective,[2] they differ from other AI systems insofar as they have proven to be effective across many different tasks and display emergent capabilities when scaled [19]. The rise of foundation models also reflects a shift in how AI systems are designed and deployed, since these models tend to be trained and released by one actor and subsequently adapted for a wide range of different applications by a plurality of other actors.

From an AI auditing perspective, foundation models pose significant challenges. For example, it is difficult to assess the risks that AI systems pose independent of the context in which they are deployed. Moreover, how to allocate responsibility between technology providers and downstream developers when harms occur remains unresolved. Taken together, the capabilities and training processes of foundation models have outpaced the development of tools and procedures to ensure that these are ethical, legal, and technically robust.[3] This implies that, while application-level audits have an important role in AI governance, they must be complemented with new forms of supervision and control.

This article addresses that gap by focusing on a subset of foundation models, namely large language models (LLMs). LLMs start from a source input, called the prompt, to generate the most likely sequences of words, code, or other data [20]. Historically, different model architectures have been used in natural language processing (NLP), including probabilistic methods [21]. However, most recent LLMs—including those we focus on in this article—are based on deep neural networks trained on a large corpus of texts. Examples of such LLMs include GPT-3 [22], GPT-4 [23], PaLM [24], LaMDA [25], Gopher [26] and OPT [27]. Once an LLM has been pre-trained, it can be adapted (with or without fine-tuning[4]) to support various applications, from spell-checking [28] to creative writing [29].

Developing LLM auditing procedures is an important and timely task for two reasons. First, LLMs pose many ethical and social challenges, including the perpetuation of harmful stereotypes, the leakage of personal data protected by privacy regulations, the spread of misinformation, plagiarism, and the misuse of copyrighted material [30-33]. In recent months, the scope of impact from these harms has been dramatically scaled by unprecedented public visibility and growing user bases of LLMs. For example, ChatGPT attracted over 100 million users just two months after its launch [34]. The urgency of addressing those challenges makes developing a capacity to audit LLMs' characteristics along different normative dimensions (such as privacy, bias, safety, etc.) a critical task in and of itself [35]. Second, LLMs can be considered proxies for other foundation models.[5] Consider CLIP [36], a vision-language model trained to predict which text caption accompanied an image, as an example. CLIP too displays emergent capabilities, can be adapted for multiple downstream applications, and faces similar governance challenges as LLMs. The same holds of text2image models such as DALL·E 2 [37]. Developing feasible and effective procedures for how to audit LLMs is therefore likely to offer transferable lessons on how to audit other foundation models and even more powerful generative systems in the future.[6]

The main contribution offered in this article is a novel blueprint for how to audit LLMs. Specifically, we propose a three-layered approach, whereby *governance audits* (of technology providers that design and disseminate LLMs), *model*

[2] Foundation models are typically based on deep neural networks and self-supervised learning, two approaches that have existed for decades [18]. That said, the rise of foundation models has been enabled by more recent developments, including: the advancement of new network architectures, like transformers [276]; the increase in compute resources and improvements in hardware capacity [277]; the availability of large scale datasets, e.g., through ImageNet [278] or CommonCrawl [279]; and the application of these increased compute resources with larger datasets for model pre-training [280].

[3] The European Commission's *Ethics Guidelines for Trustworthy AI* stipulate that AI systems should be legal, ethical, and technically robust [281]. That normative standard includes safeguards against both immediate and long-term concerns, e.g., those related to data privacy and discrimination and those related to the safety and control of highly capable and autonomous AI systems, respectively.

[4] To fine-tune LLMs for specific tasks, an additional dataset of in-domain examples can be used to adapt the final layers of a pre-trained model. In some cases, developers apply reinforcement learning (RL)—a feedback driven training paradigm whereby LLMs learn to adjust their behaviour to maximise a reward function [282]; especially reinforcement learning from human feedback (RLHF)—where the reward function is estimated based on human ratings of model outputs [50-52]. Alternatively, LLMs can be adapted to specific tasks with no additional training data and frozen weights—via in-context learning or prompt-based demonstrations [283].

[5] In some cases, the ability to utilize other modalities is integrated into single-modal LLMs: DeepMind's Flamingo model [284] fuses an LLM with visual embeddings to exploit its strong existing performance on text-based tasks.

[6] Following Jonathan Zittrain [285], we define 'generative technologies' as technologies that allow third-parties to innovate upon them without any gatekeeping. Colloquially, 'generative AI' sometimes refers to systems that can output content (images, text, audio, or code) [286], but that is not how we use the term in this article.

audits (of LLMs after pre-training but prior to their release), and *application audits* (of applications based on LLMs) complement and inform each other. Figure 1 (see Sect. 4.1) provides an overview of this three-layered approach. As we demonstrate throughout this article, many tools and methods already exist to conduct audits at each individual level. However, the key message we seek to stress is that, to provide meaningful assurance for LLMs, audits conducted on the governance, model, and application levels must be combined into a structured and coordinated procedure. Figure 2 (see Sect. 4.5) illustrates how outputs from audits on one level become inputs for which audits on other levels must account. To the best of our knowledge, our blueprint for how to audit LLMs is the first of its kind, and we hope it will inform both technology providers' and policymakers' efforts to ensure that LLMs are legal, ethical, and technically robust.

In the process of introducing and discussing our three-layered approach, the article also offers two secondary contributions. First, it makes seven claims about how LLM auditing procedures should be designed to be feasible and effective in practice. Second, it identifies the conceptual, technical, and practical limitations associated with auditing LLMs. Together, these secondary contributions lay a groundwork that other researchers and practitioners can build upon when designing new, more refined, LLM auditing procedures in the future.

Our efforts tie into an extensive research agenda and ongoing policy formation process. AI labs like Cohere, OpenAI, and AI21 have expressed interest in understanding what it means to develop LLMs responsibly [38], and DeepMind, Microsoft, and Anthropic have highlighted the need for new governance mechanisms to address the social and ethical challenges that LLMs pose [30, 39, 40]. Individual parts of our proposal (e.g., those related to model evaluation [24] and red teaming [41, 42])[7] have thus already started to be implemented across the industry, although not always in a structured manner or with full transparency. Policymakers, too, are interested in ensuring that societies benefit from LLMs while managing the associated risks. Recent examples of proposed AI regulations include the EU AI Act [43] and the US Algorithmic Accountability Act of 2022 [44]. The blueprint for auditing LLMs outlined in this article neither seeks to replace existing best practices for training and testing LLMs nor to foreclose forthcoming AI regulations. Instead, it complements them by demonstrating how governance, model, and application audits—when conducted in a structured and coordinated manner—can help ensure that LLMs are designed and deployed in ethical, legal, and technically robust ways.

A further remark is needed to narrow down this article's scope. Our three-layered approach concerns the *procedure* of LLM audits and answers questions about *what* should be audited, *when,* and according to *which criteria*. Of course, when designing a holistic auditing ecosystem, several additional considerations exist, e.g., *who* should conduct the audit and *how* to ensure post-audit action [12]. While such considerations are important, they fall outside the scope of this article. How to design an institutional ecosystem to audit LLMs is a non-trivial question that we have neither the space nor the capacity to address here. That said, the policy process required to establish an LLM auditing ecosystem will likely be gradual and involve negotiations between numerous actors, including AI labs, policymakers, and civil rights groups. For this reason, our early blueprint for how to audit LLMs is intentionally limited in scope to not forego but rather to initiate this policy formation process by eliciting stakeholder reactions.

The remainder of this article proceeds as follows: Sect. 2 highlights the ethical and social risks posed by LLMs and establishes the need to audit them. In doing so, it situates our work in relation to recent technological and societal developments. Section 3 reviews previous literature on AI auditing to identify transferable best practices, discusses the properties of LLMs that undermine existing AI auditing procedures, and derives seven claims for how LLM auditing procedures should be designed to be feasible and effective. Section 4 outlines our blueprint for how to audit LLMs, introducing a three-layered approach that combines governance, model, and application audits. The section explains in detail why these three types of audits are needed, what they entail, and the outputs they should produce. Section 5 discusses the limitations of our three-layered approach and demonstrates that any attempt to audit LLMs will face several conceptual, technical, and practical constraints. Finally, Sect. 6 concludes by discussing the implications of our findings for technology providers, policymakers, and independent auditors.

2 The need to audit LLMs

This section summarises previous research on LLMs and their ethical and social challenges. It aims to situate our work in relation to recent technological and societal developments, stress the need for auditing procedures that capture the risks LLMs pose, and address potential objections to our approach.

[7] A 'red team' is a group of people authorised to emulate an adversarial attack on a system to identify and exploit its vulnerabilities [287]. The objective of red teaming is thus to gather information that in turn can be used to improve the system's robustness.

2.1 The opportunities and risks of LLMs

Although LLMs represent a major advance in AI research, the idea of building text-processing machines is not new. Since the 1950s, NLP researchers and practitioners have been developing software that can analyse, manipulate, and generate natural language [45]. Until the 1980s, most NLP systems used logic-based rules and focused on automating the structural analysis of language needed to enable machine translation and speech recognition [46]. More recently, the advent of deep learning, advances in neural architectures such as transformers, growth in computational power and the availability of internet-scraped training data have revolutionised the field [47] by permitting the creation of LLMs that can approximate human performance on some benchmarks [48, 49]. Further advances in instruction-tuning and reinforcement learning from human feedback have improved model capabilities to predict user intent and respond to natural language requests [50-52].

LLMs' core training task is to produce the most likely continuation of a text sequence [53]. Consequently, LLMs can be used to recognise, summarise, translate, and generate texts, with near human-like performance on some tasks [54]. Exactly when a language model becomes 'large' is a matter of debate—referring to either more trainable parameters [55], a larger training corpus [56] or a combination of these. For our purposes, it is sufficient to note that LLMs are highly adaptable to various downstream applications, requiring fewer in-domain labelled examples than traditional deep learning systems [57]. This means that LLMs can more easily be adapted for specific tasks, such as diagnosing medical conditions [58], generating code [59, 60] and translating languages [61]. Previous research has demonstrated that LLMs can perform well on a task with few-shot or zero-shot reasoning [22, 62].[8] Moreover, a scaling law has been identified whereby the training error of an LLM falls off as a power of training set size, model size or both [63]. Simply scaling the model can thus result in emergent gains on a wide array of tasks [64], though those gains are non-uniform, especially for complex mathematical or logical reasoning domains [26]. Finally, while some pre-trained models are protected by paywalls or siloed within companies, many LLMs are accessible via open-source libraries such as HuggingFace, democratising the gains from deep language modelling and allowing non-experts to use it in their applications [65].

Alongside such opportunities, however, the use of LLMs is coupled with ethical challenges [31, 32]. As recent controversies surrounding ChatGPT [66] have shown, LLMs are prone to give biased or incorrect answers to user queries [67]. More generally, a recent article by Weidinger et al. [30] suggests that the risks associated with LLM include the following:

(1) *Discrimination.* LLMs can introduce representational and allocational harms by perpetuating social stereotypes and biases;
(2) *Information hazards.* LLMs may compromise privacy by leaking private information and inferring sensitive information;
(3) *Misinformation hazards.* LLMs producing misleading information can lead to less well-informed users and erode trust in shared information;
(4) *Malicious use.* LLMs can be co-opted by users with bad intent, e.g., to generate personalised scams or large-scale fraud;
(5) *Human–computer interaction harms.* Users may overestimate the capabilities of LLMs that appear human-like and use them in unsafe ways; and
(6) *Automation and environmental harms.* Training and operating LLMs require lots of computing power, incurring high environmental costs.

Each of these risk areas constitutes a vast and complex field of research. Providing a comprehensive overview of each field's nuances is beyond this paper's scope. Instead, we take Weidinger et al.'s summary of the ethical and social risks associated with LLMs as a starting point for pragmatic problem-solving.

2.2 The governance gap

From a governance perspective, LLMs pose both methodological and normative challenges. As previously mentioned, foundation models—like LLMs—are typically developed and adopted in two stages. Firstly, a model is pre-trained using self-supervised learning on a large, unstructured text corpus scraped from the internet. Pre-training captures the general language representations required for many tasks without explicitly labelled data. Secondly, the weights or behaviours of this pre-trained model can be adapted on a far smaller dataset of labelled, task-specific, examples.[9] That makes it methodologically difficult to assess LLMs independent of the context in which they will be deployed [18].

Furthermore, although performance is predictable at a general level, performance on specific tasks, or at scale, can be unpredictable [40]. Crucially, even well-functioning

[8] An LLM is considered a 'zero-shot' reasoner if employed for a completely unseen task and a 'few-shot' reasoner if only a small sample of demonstrations are given for a previously unseen task.

[9] The term 'adapted' here encompasses multiple existing methods for eliciting specific model behaviours, including fine-tuning, reinforcement learning with human feedback and in-context learning.

LLMs force AI labs and policymakers to face hard questions, such as who should have access to these technologies and for which purposes [68]. Of course, the challenges posed by LLMs are not necessarily distinct from those associated with classical NLP or other ML-based systems. However, LLMs' widespread use and generality make those challenges deserving of urgent attention. For all these reasons, analysing LLMs from ethical perspectives requires innovation in risk assessment tools, benchmarks, and frameworks [69].

Several governance mechanisms designed to ensure that LLMs are legal, ethical, and safe have been proposed or piloted [70]. Some are technically oriented, including the pre-processing of training data, the fine-tuning of LLMs on data with desired properties, and procedures to test the model at scale pre-deployment [42, 69]. Others seek to address the ethical and social risks associated with LLMs through sociotechnical mitigation strategies, e.g., creating more diverse developer teams [71], human-in-the-loop protocols [72] and qualitative evaluation tools based on ethnographic methods [73]. Yet others seek to ensure transparency in AI development processes, e.g., through a structured use of model cards [74, 75], datasheets [76], system cards [77], and the watermarking of system outputs [78].[10]

To summarise, while LLMs have shown impressive performance across a wide range of tasks, they also pose significant ethical and social risks. Therefore, the question of how LLMs should be governed has attracted much attention, with proposals ranging from structured access protocols designed to prevent malicious use [68] to hard regulation prohibiting the deployment of LLMs for specific purposes [79]. However, the effectiveness and feasibility of these governance mechanisms have yet to be substantiated by empirical research. Moreover, given the multiplicity and complexity of the ethical and social risks associated with LLMs, we anticipate that policy responses will need to be multifaceted and incorporate several complementary governance mechanisms. As of now, technology providers and policymakers have only started experimenting with different governance mechanisms, and how LLMs should be governed remains an open question [80].

2.3 Calls for audits

Against the backdrop of the technological and regulatory landscape surveyed in this section, auditing should be understood as one of several governance mechanisms different stakeholders can employ to ensure and demonstrate that LLMs are legal, ethical, and technically robust. It is important to stress that auditing LLMs is not a hypothetical idea but a tangible policy option that has been proposed by researchers, technology providers, and policymakers alike. For instance, when coining the term foundation models, Bommasani et al. [18] suggested that 'such models should be subject to rigorous testing and *auditing* procedures'. Moreover, in an open letter concerning the risks associated with LLMs and other foundation models, OpenAI's CEO Sam Altman stated that 'it's important that efforts like ours submit to *independent audits* before releasing new systems' [81]. Finally, the European Commission is considering classifying LLMs as 'high-risk AI systems' [82].[11] This would imply that technology providers designing LLMs have to undergo 'conformity assessments with the involvement of an independent third-party', i.e., audits by another name [83].

Despite widespread calls for LLM auditing, central questions concerning *how* LLMs can and should be audited have yet to be systematically explored. This article addresses that gap by outlining a procedure for auditing LLMs. The main argument we advance can be summarised as follows. What auditing means varies between different academic disciplines and industry contexts [84]. However, three strands of auditing research and practice are particularly relevant with respect to ensuring good governance of LLMs. The first stems from IT audits, whereby auditors assess the adequacy of technology providers' software development processes and quality management procedures [85]. The second strand stems from model testing and verification within the computer sciences, whereby auditors assess the properties of different computational models [86]. The third strand stems from product certification procedures, whereby auditors test consumer goods for legal compliance and technical safety before they go to market [87]. As we argue throughout this paper, it is necessary to combine auditing tools and procedural best practices from each of these three strands to identify and manage the social and ethical risks LLMs pose. Therefore, our blueprint for auditing LLMs combines *governance audits* of technology providers, *model audits* of LLMs, and *application audits* of downstream products and services built on top of LLMs. The details of this 'three-layered approach' are outlined in Sect. 4.

2.4 Addressing initial objections

Before proceeding any further, it is useful to consider some reasonable objections to the prospect of auditing LLMs—as well as potential responses to these objections. First, one may argue that there is no need to audit LLMs per se and

[10] A watermark is a hidden pattern in a text that is imperceptible to humans but makes it algorithmically identifiable as synthetic.

[11] It is still uncertain how the EU AI Act should be interpreted. The current formulation states that LLMs that may be used for high-risk applications should be considered high-risk [288].

that auditing procedures should be established at the application level instead. Although audits on the application level are important, the objection presents a false dichotomy: quality and accountability mechanisms can and should be established at different stages of supply chains. Moreover, while some risks can only be addressed at the application level, others are best managed upstream. It is true that many factors, including some beyond the technology provider's control, determine whether a specific technological artefact causes harm [88]. However, technology providers are still responsible for taking proportional precautions regarding reasonably foreseeable risks during the product life cycle stages that they do control. For this reason, we propose that application audits should be complemented with governance audits of the organisations that develop LLMs. The same logic underpins the EU's AI liability directive [89]. Our proposal is thereby compatible with the emerging European AI regulations.

Second, identifying and mitigating all LLM-related risks at the technology level may not be possible. As we explain in Sect. 5, this is partly because different normative values may conflict and require trade-offs [90-92]. Using individuals' data, for example, may permit improved personalisation of language models, but compromise privacy [93]. Moreover, concepts like 'fairness' or 'transparency' hide deep normative disagreements [94]. Different definitions of fairness (like demographic parity and counterfactual fairness) are mutually exclusive [95-97], and prioritising between competing definitions remains a political question. However, while audits cannot ensure that LLMs are 'ethical' in any universal sense, they nevertheless contribute to good governance in several ways. For example, audits can help technology providers identify risks and potentially prevent harm, shape the continuous (re-design) of LLMs, and inform public discourse concerning tech policy. Bringing all this together, our blueprint for how to audit LLMs focuses on making implicit choices and tensions visible, giving voice to different stakeholders, and generating resolutions that—even when imperfect—are, at least, more explicit and publicly defensible [98].

Third, one may contend that designing LLM auditing procedures is difficult. We agree and would add that this difficulty has both practical and conceptual components. Different stages in the software development life cycle (including curating training data and the pre-training/fine-tuning of model weights) overlap in messy and iterative ways [99]. For example, open-source LLMs are continuously re-trained and re-uploaded on collaborative platforms (like HuggingFace) post-release. That creates practical problems concerning when and where audits should be mandated. Yet the conceptual challenges run even more deeply. For instance, what constitutes disinformation and hate speech are contested questions [100]. Despite widespread agreement that LLMs should be 'truthful' and 'fair', such notions are hard to operationalise. Because there exists no universal condition of validity that applies equally to all kinds of utterances [101], it is hard to establish a normative baseline against which LLMs can be audited.

However, these difficulties are not reasons for abstaining from developing LLM auditing procedures. Instead, they are healthy reminders that it cannot be assumed that one single auditing procedure will capture all LLM-related ethical risks or be equally effective in all contexts [102]. The insufficiency and limited nature of auditing as a governance mechanism is not an argument against its complementary usefulness. With those caveats highlighted, we now review previous work on AI auditing. The aim of the next section is thus to explore the merits and limitations of existing AI auditing procedures when applied to LLMs and, ultimately, identify transferable best practices.

3 The merits and limits of existing AI auditing procedures

In this section, we provide an overview of previous work.[12] In doing so, we introduce auditing as an AI governance mechanism, highlight the properties of LLMs that undermine the feasibility and effectiveness of existing AI auditing procedures, and derive and defend seven claims about how LLM auditing procedures should be designed. Taken together, this section provides the theoretical justification for the LLM auditing blueprint outlined in Sect. 4.

3.1 AI auditing

In the broadest sense, auditing refers to an independent examination of any entity, conducted with a view to express an opinion thereon [103]. Auditing can be conceived as a governance mechanism because it can be used to monitor conduct and performance [104] and has a long history of promoting procedural regularity and transparency in areas like financial accounting and worker safety [105]. The idea behind AI auditing is thus simple: just like financial transactions can be audited for correctness, completeness, and legality, so can the design and use of AI systems be audited for technical robustness, legal compliance, or adherence with pre-defined ethics principles.

AI auditing is a relatively recent field of study, sparked in 2014 by Sandvig et al.'s article *Auditing Algorithms* [1]. However, auditing intersects with almost every aspect of AI governance, from the documentation of design procedures to model testing and verification [106]. AI auditing is thus both a multifaceted practice and a multidisciplinary field of

[12] See Appendix 1 for the methodology used to conduct this literature review.

research, harbouring contributions from computer science [107, 108], law [109, 110], media and communication studies [1, 111], and organisation studies [112, 113].

Different researchers have defined AI auditing in different ways. For example, it is possible to distinguish between narrow and broad conceptions of AI auditing. The former is impact-oriented and focuses on probing and assessing the outputs of AI systems for different input data [114]. The latter is process-oriented and focuses on assessing the adequacy of technology providers' software development processes and quality management systems [115]. This article takes the broad perspective, defining AI auditing as a systematic and independent process of obtaining and evaluating evidence regarding an entity's actions or properties and communicating the results of that evaluation to relevant stakeholders. Note that the entity in question, i.e., the audit's subject, can be either an AI system, an organisation, a process, or any combination thereof [116].

Different actors can employ AI auditing for different purposes [117]. In some cases, policymakers mandate audits to ensure that AI systems used within their jurisdiction meet specific legal standards. For example, New York City's AI Audit Law (NYC Local Law 144) requires independent auditing of companies utilising AI systems to inform employment-related decisions [118]. In other cases, technology providers commission AI audits to mitigate technology-related risks, calling on professional services firms like PwC, Deloitte, KPMG, and EY [119-122]. In yet other cases, other stakeholders conduct AI audits to inform citizens about the conduct of specific companies.[13]

The key takeaway from this brief overview is that while AI auditing is a widespread practice, both the design and purpose of different AI auditing procedures vary. Moreover, procedures to audit LLMs and other foundation models have yet to be developed. Therefore, it is useful to consider the merits and limitations of existing AI auditing procedures when applied to LLMs.

3.2 Seven claims about auditing LLMs

As demonstrated above, a wide range of AI auditing procedures have already been developed.[14] However, not all auditing procedures are equally effective in handling the risks posed by LLMs. Nor are they equally likely to be implemented, due to factors including technical limitations, institutional access, and administrative costs [3]. In what follows, we discuss some key distinctions that inform the design of auditing procedures and defend seven claims about making such designs feasible and effective for LLMs.

To start with, it is useful to distinguish between *compliance audits* and *risk audits*. The former compares an entity's actions or properties to predefined standards or regulations. The latter asks open-ended questions about how a system works to identify and control risks. When conducting risk audits of LLMs, auditors can draw on well-established procedures, including standards for AI risk management [123, 124] and guidance on how to assess and evaluate AI systems [112, 125-129]. In contrast, compliance audits require a normative baseline against which AI systems can be evaluated. However, LLM research is a quickly developing field in which standards and regulations have yet to emerge. Moreover, the fact that LLMs are adaptable to many downstream applications [40] undermines the feasibility of auditing procedures designed to ensure compliance with sector-specific norms and regulations. This leads us to our first claim:

Claim 1 *AI auditing procedures focusing on compliance alone are unlikely to provide adequate assurance for LLMs.*

Our blueprint for how to audit LLMs outlined in Sect. 4 accounts for Claim 1 by incorporating elements of both risk audits (at governance and model levels) and compliance audits (at the application level).

Further, it is useful to distinguish between *external* and *internal audits*. The former is conducted by independent third-parties and the latter by an internal function reporting directly to its board [130]. External audits help address concerns regarding accuracy in self-reporting [1], so they typically underpin formal certification procedures [131]. However, they are constrained by limited access to internal processes [9]. For internal audits, the inverse is true: while constituting an essential step towards informed model design decisions [132], they run an increased risk of collusion between the auditor and the auditee [133]. Moreover, without third-party accountability, decision-makers may ignore audit recommendations that threaten their business interests [134]. The risks stemming from misaligned incentives are especially stark for technologies with rapidly increasing capabilities and for companies facing strong competitive pressures [135]. Both conditions apply to LLMs, undermining the ability of internal auditing procedures to provide meaningful assurance in this space. This observation, combined with the need to manage the social and ethical risks posed by LLMs surveyed in Sect. 2, leads us to assert that:

[13] AI auditing procedures have not only been developed by academic researchers and private companies but also by non-profit organisations like ForHumanity [302] and the Algorithmic Justice League [289].

[14] For more comprehensive overviews of available AI auditing tools and procedures, see [6, 114, 117].

Claim 2 *External audits are required to ensure that LLMs are ethical, legal, and technically robust, as well as to hold technology providers accountable in case of irregularities of incidents.*

As we explain in Sect. 4, each step in our blueprint for how to audit LLMs should be conducted by independent third-party auditors. However, external audits come with their own challenges, including how to access information that is protected by privacy or IP rights [12, 136]. This is especially challenging in the case of LLMs since some are only accessible via an application programming interface (API) and others are not published at all. Determining the auditor's level of access is thus an integral part of designing LLM auditing procedures.

Koshiyama et al. [10] proposed a typology that distinguishes between different access levels. At lower levels, auditors have no direct access to the model but base their evaluations on publicly available information about the development process. At middle levels, auditors have access to the computational model itself, meaning they can manipulate its parameters and review its task objectives. At higher levels, auditors have access equivalent to the system developer to all the details encompassing a system, i.e., full access to organisational processes, actual input and training data, and information about how and why the system was initially created. In Sect. 4, we use this typology to indicate the level of access auditors need to conduct audits at the governance, model, and application levels.

The question about access leads us to a further distinction made in the AI auditing literature, i.e., between *adversarial* and *collaborative* audits. Adversarial audits are conducted by independent actors to assess the properties or impact an AI system has—without privileged access to its source code or technical design specifications [1, 114]. Collaborative audits see technology providers and external auditors working together to assess and improve the process that shapes future AI systems' design and safeguards [115, 116]. While the former primarily aims to expose harms, the latter seeks to provide assurance. Previous research has shown that audits are most effective when technology providers and independent auditors collaborate towards the common goal of identifying and managing risks [11]. This implies that:

Claim 3 *To be feasible and effective in practice, procedures to audit LLM require active collaboration between technology providers and independent auditors.*

Accounting for Claim 3, this article focuses on collaborative audits. All steps in our three-layered approach outlined in Sect. 4 demand that technology providers provide external auditors with the access they need and proactively feed their own know-how into the process. After all, evaluating LLMs requires resources and technical expertise that technology providers are best positioned to provide.

Moving on, it is also useful to distinguish between *governance audits* and *technology audits*. The former focus on the organisation designing or deploying AI systems and include assessments of software development and quality management processes, incentive structures, and the allocation of roles and responsibilities [85]. The latter focus on assessing a technical system's properties, e.g., reviewing the model architecture, checking its consistency with predefined specifications, or repeatedly querying an algorithm to understand its workings and potential impact [114]. Some LLM-related risks can be identified and mitigated at the application level. However, other issues are best addressed upstream, e.g., those concerning the sourcing of training data. This implies that, to be feasible and effective:

Claim 4 *Auditing procedures designed to assess and mitigate the risks posed by LLMs must include elements of both governance and technology audits.*

Our blueprint for how to audit LLMs satisfies this claim in the following way. The governance audits we propose aim to assess the processes whereby LLMs are designed and disseminated, the model audits focus on assessing the technical properties of pre-trained LLMs, and the application audits focus on assessing the technical properties of applications built on top of LLMs.

However, both governance audits and technology audits have limitations. During governance audits, for example, it is not possible to anticipate upfront all the risks that emerge as AI systems interact with complex environments over time [102, 137]. Further, not all ethical tensions stem from technology design alone, as some are intrinsic to specific tasks or applications [138]. While these limitations of governance audits are well-known, LLMs introduce new challenges for technology audits, which have historically focused on assessing systems designed to fill specific functions in well-defined contexts, e.g., improving image analysis in radiology [139] or detecting corporate fraud [140]. Because LLMs enable many downstream applications, traditional auditing procedures are not equipped to capture the full range social and ethical risks they pose. While existing best practices in governance auditing appear applicable to organisations designing or deploying LLMs, that is not true for technology audits. In short:

Claim 5 *The methodological design of technology audits will require significant modifications to identify and assess LLM-related risks.*

As mentioned above, our blueprint for how to audit LLMs incorporates elements of technology audits on both the model and the application levels. To understand why that is necessary to identify and mitigate the ethical risks posed by LLMs, we must first distinguish between different types of technology audits.

Previous work on technology audits distinguish between *functionality*, *model*, and *impact audits* [141]. Functionality audits focus on the rationale underpinning AI systems by asking questions about intentionality, e.g., what is this system's purpose [142]? Model audits review the system's decision-making logic. For symbolic AI systems,[15] that entails reviewing the source code. For sub-symbolic AI systems, including LLMs, it entails asking how the model was designed, what data it was trained on, and how it performs on different benchmarks. Finally, impact audits investigate the types, severity, and prevalence of effects from an AI system's outputs on individuals, groups, and the environment [143]. These approaches are not mutually exclusive but rather highly complementary [116]. Still, technology providers that design and disseminate LLMs have limited information about the future deployment of their systems by downstream developers and end-users. This leads us to our sixth claim:

Claim 6 *Model audits will play a key role in identifying and communicating LLMs' limitations, thereby informing system redesign, and mitigating downstream harm.*

This claim constitutes a key justification for the three-layered approach to LLM auditing proposed in this article. As highlighted in Sect. 4, governance audits and application audits are both well-established practices in systems engineering and software development. Hence, it is precisely by adding structured and independent audits on the model level that our blueprint for auditing LLMs complements and enhances existing governance structures.

Finally, within technology audits, it is important to distinguish between ex-ante and ex-post audits, which take place before and after a system is deployed, respectively. The former can identify and prevent some harms before they occur while informing downstream users about the model's appropriate, intended applications. Considerable literature already exists within computer science on techniques such as red teaming [41, 42], model fooling [144], functional testing [145] and template-based stress-testing [146], which all play important roles during technology audits of LLMs. However, ex-ante audits cannot fully capture all the risks associated with systems that continue to 'learn' by updating their internal decision-making logic [147].[16] This limitation applies to all learning systems but is particularly relevant for LLMs that display emergent capabilities [148].[17] Ex-post audits can be divided into *snapshot audits* (which occur once or on regular occasions) and *continuous audits* (which monitor performance over time). Most existing AI auditing procedures are snapshots.[18] Like ex-ante audits, however, snapshots are unable to provide meaningful assurance regarding LLMs as they display emergent capabilities and, in some cases, can learn as they are fed new data. This leads to our final claim:

Claim 7 *LLM auditing procedures must include elements of continuous ex-post monitoring to meet their regulatory objectives.*

In our blueprint, continuous ex-post monitoring is one of the activities conducted at the application level. However, as detailed in Sect. 4.5, audits on the different levels are strongly interconnected. For example, continuous monitoring of LLM-based applications presupposes that technology providers have established ex-post monitoring plans—which can only be verified by audits at the governance level. Invertedly, technology providers rely on feedback from audits at the application level to continue improving their software development and quality management procedures.

To summarise, much can be learned from existing AI auditing procedures. However, LLMs display several properties that undermine the feasibility of such procedures. Specifically, LLMs are adaptable to a wide range of downstream applications, display emergent capabilities, and can, in some cases, continue to learn over time. As this section has shown, that means that neither functionality audits (which hinge on the evaluation of the purpose of a specific application) nor impact audits (which hinge on the ability to observe a specific system's actual impact) alone can provide meaningful assurance against the social and ethical risks LLMs pose. It also means that ex-ante audits must be complemented by continuous post-market monitoring of outputs from LLM-based applications.

In this section, we have built on these and other insights to derive and defend seven claims about how auditing procedures should be designed to account for the governance challenges LLMs pose. These seven claims provided our

[15] Symbolic AI systems are based on explicit methods like first-order logic and decision trees. Sub-symbolic systems rely on establishing correlations through statistical methods like Bayesian learning and back-propagation [290].

[16] In their unfrozen states, all LLMs can learn as they are fed new data. However, once a model has been 'fixed', it does not update and simply uses new input data to make predictions.

[17] Emergence implies that an entity can have properties its parts do not individually possess, and that randomness can generate orderly structures [291].

[18] The post-market monitoring mandated by the proposed EU AI Act [43] is a rare example of continuous auditing.

starting point when designing the three-layered approach for auditing LLMs that will be outlined in Sect. 4. However, we maintain that these claims are more general and could serve as guardrails for other attempts to design auditing procedures for all foundation models.

4 Auditing LLMs: a three-layered approach

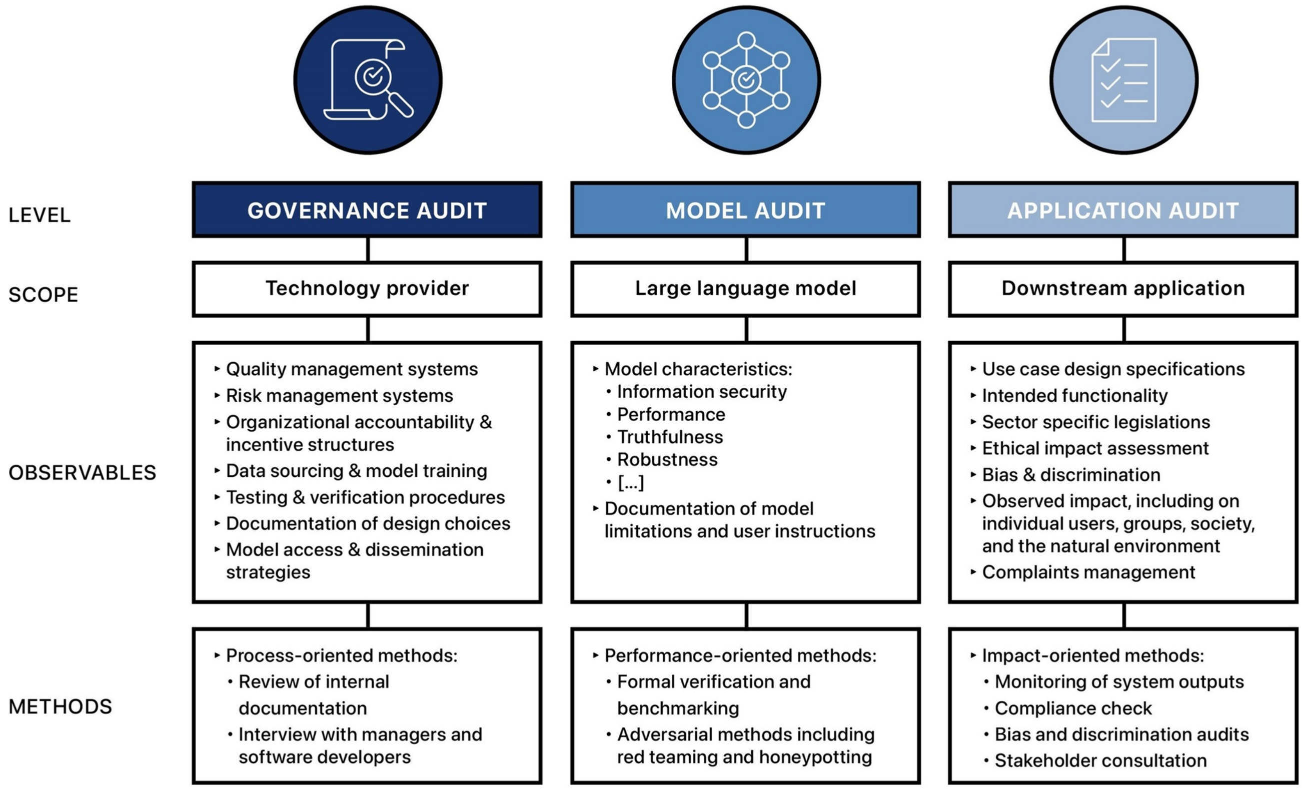
This section offers a blueprint for auditing LLMs that satisfies the seven claims in Sect. 3 about how to structure such procedures. While there are many ways to do that, our proposal focuses on a limited set of activities that are (i) jointly sufficient to identify LLM-related risks, (ii) practically feasible to implement, and (iii) have a justifiable cost–benefit ratio. The result is the three-layered approach outlined below.

4.1 A blueprint for LLM auditing

Audits should focus on three levels. First, technology providers developing LLMs should undergo *governance audits* that assess their organisational procedures, accountability structures and quality management systems. Second, LLMs should undergo *model audits*, assessing their capabilities and limitations after initial training but before adaptation and deployment in specific applications. Third, downstream applications using LLMs should undergo continuous *application audits* that assess the ethical alignment and legal compliance of their intended functions and their impact over time. Figure 1 illustrates the logic of our approach.

Some clarifications are needed to flesh out our blueprint. To begin with, governance, model and application audits only provide effective assurance when coordinated. This is because the affordances and limitations of audits conducted at the three levels differ in ways that make them critically complementary. For example, as Sect. 3 showed, LLM audits must include elements of both process- and performance-oriented auditing (Claim 4). In our three-layered approach, the governance audits are process-oriented, whereas the model and application audits are performance-oriented. Moreover, feasible and effective LLM auditing procedures must include aspects of continuous, ex-post assessments (Claim 7). In our blueprint, these elements are incorporated at the application level. But this is just two examples. As we discuss what governance, model and applications audits entail in this section, we also make highlight how they, when combined, satisfies all seven claims listed in Sect. 3.

While the three types of audits included in our blueprint are individually necessary, their boundaries overlap and can



LEVEL	GOVERNANCE AUDIT	MODEL AUDIT	APPLICATION AUDIT
SCOPE	Technology provider	Large language model	Downstream application
OBSERVABLES	‣ Quality management systems ‣ Risk management systems ‣ Organizational accountability & incentive structures ‣ Data sourcing & model training ‣ Testing & verification procedures ‣ Documentation of design choices ‣ Model access & dissemination strategies	‣ Model characteristics: • Information security • Performance • Truthfulness • Robustness • [...] ‣ Documentation of model limitations and user instructions	‣ Use case design specifications ‣ Intended functionality ‣ Sector specific legislations ‣ Ethical impact assessment ‣ Bias & discrimination ‣ Observed impact, including on individual users, groups, society, and the natural environment ‣ Complaints management
METHODS	‣ Process-oriented methods: • Review of internal documentation • Interview with managers and software developers	‣ Performance-oriented methods: • Formal verification and benchmarking • Adversarial methods including red teaming and honeypotting	‣ Impact-oriented methods: • Monitoring of system outputs • Compliance check • Bias and discrimination audits • Stakeholder consultation

Fig. 1 Blueprint for how to audit LLMs: A three-layered approach

be drawn in multiple ways. For example, the collection and pre-processing of training data ties into software development practices. Hence, reviewing organisational procedures for obtaining and curating training data is legitimate during holistic governance audits. However, the characteristics LLMs display during model audits may also reflect biases in their training data [149, 150].[19] Reviewing such data is, therefore, often necessary during the model audits too [151, 152]. Nevertheless, the conceptual distinction between governance, model and application audits remains useful when identifying varied risks that LLMs pose.

It is theoretically possible to add further layers to our blueprint. For example, downstream developers could also be made subject to process-oriented governance audits. But such audits would be difficult to implement, given that many decentralised actors build applications on top of LLMs. The combination of governance, model, and application audits, we argue, strikes a balance between covering a sufficiently large part of the development and deployment lifecycle to identify LLM-related risks, on the one hand, and being practically feasible to implement, on the other. Regardless of how many layers are included, however, the success of our blueprint relies on responsible actors at each level who actively want to or are incentivised to ensure good governance.

Finally, to provide meaningful assurance, audits on all three levels should be external (Claim 2) yet collaborative (Claim 3). In practice, this implies that independent third parties not only seek to verify claims made by technology providers but also work together with them to identify and mitigate risks and shape the design of future LLMs. As mentioned in the introduction, the question of who should conduct the audits falls outside the scope of this article. That said, reasonable concerns about how independent collaborative audits really are can be raised regardless of who is conducting the audit. In Sect. 5, we discuss this and other limitations.

With those clarifications in mind, we will now present the details of our three-layered approach. The following three subsections discuss governance, model, and application audits respectively, focusing on why each is needed, what each entails, and what outputs each should produce.

4.2 Governance audits

Technology providers working on LLMs should undergo governance audits that assess their organisational procedures, incentive structures, and management systems. Overwhelming evidence shows that such features influence the design and deployment of technologies [4]. Moreover, research has demonstrated that risk-mitigation strategies work best when adopted transparently, consistently, and with executive-level support [153, 154]. Technology providers are responsible for identifying the risks associated with their LLMs and are uniquely well-positioned to manage some of those risks. Therefore, it is crucial that their organisational procedures and governance structures are adequate.

Governance audits have a long history in areas like IT governance [85, 155, 156] and systems and safety engineering [157-159]. Tasks include assessing internal governance structures, product development processes and quality management systems [115] to promote transparency and procedural regularity, ensure that appropriate risk management systems are in place [160], and spark deliberation regarding ethical and social implications throughout the software development lifecycle. Governance audits can also improve accountability, e.g., publicising their results prevents companies from covering up undesirable outcomes and incentivises better behaviour [136]. Thus defined, governance audits incorporate elements of both compliance audits, regarding completeness and transparency of documentation, and risk audits, regarding the adequacy of the risk management system (Claim 1).

Specifically, we argue that governance audits of LLM providers should focus on three tasks:[20]

(1) *Reviewing the adequacy of organisational governance structures* to ensure that model development processes follow best practices and that quality management systems can capture LLM-specific risks. While technology providers have in-house quality management experts, confirmation bias may prevent them from recognising critical flaws; involving external auditors addresses that issue [161]. Nevertheless, governance audits are most effective when auditors and technology providers collaborate to identify risks [162]. Therefore, it is important to distinguish accountability from blame at this stage of an audit.

(2) *Creating an audit trail of the LLM development process* to provide chronological documentary evidence of the development of an LLM's capabilities, including information about its intended purpose, design specifications and choices, as well as how it was trained and tested through the generation of model cards [74] and system cards [77].[21] This includes the structured use

[19] The link between model characteristics and biases in the training data can sometimes be counterintuitive [292]. In some cases, biased datasets can help models recognise bias and steer away from it with the help of reinforcement learning and human feedback [293].

[20] Governance audits could examine many tasks, and prioritization may vary depending on the sector and jurisdiction. Hence, the three tasks we propose are merely a minimum baseline.

[21] Meta AI's detailed notes on the OPT model provide an exemplar of model training documentation [294].

of datasheets [76] to document how the datasets used to train and validate LLMs were sources, labelled, and curated. The creation of such audit trails serves several related purposes. Stipulating design specifications upfront facilitates checking system adherence to jurisdictional requirements downstream [157]. Moreover, information concerning intended use cases should inform licensing agreements with downstream developers [163], thereby restricting the potential for harm through malicious use. Finally, requiring providers to document and justify their design choices sparks ethical deliberation by making trade-offs explicit.

(3) *Mapping roles and responsibilities within organisations that design LLMs* to facilitate the allocation of accountability for system failures. LLMs' adaptability downstream does not exculpate technology providers from all responsibility. Some risks are 'reasonably foreseeable'. In the adjacent field of machine learning (ML) image recognition, a study found that commercial gender classification systems were less accurate for darker-skinned females than lighter-skin males [15]. After the release of these findings, all technology providers speedily improved the accuracy of their models, suggesting that the problem was not intrinsic, but resulted from inadequate risk management. Mapping the roles and responsibilities of different stakeholders improves accountability and increases the likelihood of impact assessments being structured rather than ad-hoc, thus helping identify and mitigate harms proactively.

To conduct these three tasks, auditors primarily require what Koshiyama et al. [10] refer to as white-box auditing. This is the highest level of access and suggests that the auditor knows how and why an LLM was developed. In practice, it implies privileged access to facilities, documentation, and personnel, which is standard practice in governance audits in other fields. For example, IT auditors have full access to material and reports related to operational processes and performance metrics [85]. It also implies access to the input data, learning procedures, and task objectives used to train LLMs. White-box auditing requires that nondisclosure and data-sharing agreements are in place, which adds to the logistical burden of governance audits. However, granting such a high level of access is especially important from an AI safety perspective because, in addition to auditing LLMs before market deployment, governance audits should also evaluate organisational safeguards concerning high-risk projects that providers may prefer not to discuss publicly.

The results of governance audits should be provided in formats tailored to different audiences. The primary audience is the management and directors of the LLM provider. Auditors should provide a full report that directly and transparently lists and discusses the vulnerabilities of existing governance structures. Such reports may recommend actions, but taking actions remains the provider's responsibility. Usually, such audit reports are not made public. However, some evidence obtained during governance audits can be curated for two secondary audiences: law enforcers and developers of downstream applications. In some jurisdictions, hard legislation may demand that technology providers follow specific requirements. For instance, the proposed EU AI Act required providers to register high-risk AI systems with a centralised database [43] or implement a risk management system [164]. In such cases, reports from independent governance audits can help providers demonstrate adherence to legislation. Reports from governance audits also help developers of downstream applications to understand an LLM's intended purpose, risks, and limitations.

Before concluding this discussion, it is useful to reflect on how governance audits contribute to relieving some of the social and ethical risks LLMs pose. As mentioned in Sect. 2, Weidinger et al. [30] listed six broad risk areas: discrimination, information hazards, misinformation hazards, malicious use, human–computer interaction harm, and automation and environmental harms. Governance audits address some of these directly. By assessing the adequacy of the governance structures surrounding LLMs, including licencing agreements [163] and structured access protocols [68], governance audits help reduce the risk of malicious use. Further, some information hazards stem from the possibility of extracting sensitive information from LLMs via adversarial attacks [165]. By reviewing the process whereby training datasets were sourced, labelled, and curated, as well as the strategies and techniques used during the model training process—such as differential privacy [166] or secure federated learning [167]—governance audits can minimise the risk of LLMs leaking sensitive information. However, for most of the risk areas listed by Weidinger et al. [30], governance audits have only an indirect impact insofar as they contribute to transparency about the limitations and intended purposes of LLMs. Hence, risks areas like discrimination, misinformation hazards, and human–computer interaction harms are better addressed by model and application audits.

4.3 Model audits

Before deployment, LLMs should be subject to model audits that assess their capabilities and limitations (Claim 6). Model audits share some features with governance audits. For instance, both happen before an LLM is adapted for specific applications. However, model audits do not focus on organisational procedures but on LLMs' capabilities and characteristics. Specifically, they should identify an LLM's limitations to (i) inform the continuous redesign the system, and (ii) communicate its capabilities and limitations to external stakeholders. These two tasks use similar methodologies, but they target different audiences.

The first task—limitation identification—aims primarily to support organisations that develop LLMs with benchmarks or other data points that inform internal model redesigning and retraining efforts [168]. Model audits' results should also inform API license agreements, helping prevent applications in unintended use cases [163] and restricting the distribution of dangerous capabilities [68]. The second task—communicating capabilities and limitations—aims to inform the design of specific applications built on top of LLMs by downstream developers. Such communication can take different forms, e.g., *interactive model cards* [169], specific language model risk cards [75], and *information about the initial training dataset* [170, 171], to help downstream developers adapt the model appropriately.

In Sect. 3, we argued that the way technology audits are being conducted requires modifications to address the governance challenges associated with LLMs (Claim 5). In what follows, we demonstrate that evaluating an LLM's characteristics independent of an intended use case is challenging but not impossible.[22] To do so, auditors can use two distinct approaches. The first involves identifying and assessing intrinsic characteristics. For example, the training dataset can be assessed for completeness and consistency without reference to specific use cases [112]. However, it is often expensive and technically challenging to interrogate large datasets [172]. The second involves employing an indirect approach that tests the model across multiple potential downstream use cases, links the results to different characteristics, and assesses the aggregated results using different weighting techniques. That second approach may prove more fruitful when assessing an LLM's performance.

Nevertheless, selecting the characteristics to focus on during model audits remains challenging. Given such audits' purpose, we recommend examining characteristics that are (i) *socially and ethically relevant*, i.e., can be directly linked to the social and ethical risks posed by LLMs; (ii) *predictably transferable*, i.e., impact the nature of downstream applications; and (iii) *meaningfully operationalisable,* i.e., can be assessed with the available tools and methods.

Keeping those criteria in mind, we posit that model audits should focus on (at least) the performance, robustness, information security and truthfulness of LLMs. As other characteristics may meet the three criteria listed above, those four characteristics are just examples highlighting the role of model audits in our three-layered approach. The list of relevant model characteristics can be amended as required when developing specific auditing procedures. With those caveats out of the way, we now proceed to discuss how four example characteristics can be assessed during model audits:

(1) *Performance,* i.e., how well the LLM functions on various tasks. Standardised benchmarks can help assess an LLM's performance by comparing it to a human baseline. For example, *GLUE* [173] aggregates LLM performance across multiple tasks into a single reportable metric. Such benchmarks have been criticised for overestimating performance over a narrow set of capabilities and quickly becoming saturated, i.e., rapidly converging on the performance of non-expert humans, leaving limited space for valuable comparisons. Therefore, it is crucial to evaluate LLMs' performance against many tasks or benchmarks, and sophisticated tools and methods have been proposed for that purpose, including *SuperGLUE* [49], which is more challenging and 'harder to game' with narrow LLM capabilities, and *BIG-bench* [64], which can assess LLM's performance on tasks that appear beyond their current capabilities. These benchmarks are particularly relevant for model audits because they were primarily developed to evaluate pre-trained models, without task-specific fine-tuning.

(2) *Robustness,* i.e., how well the model reacts to unexpected prompts or edge cases. In ML, robustness indicates how well an algorithm performs when faced with new, potentially unexpected (i.e., out-of-domain) input data. LLMs lacking robustness introduce, at least, two distinct risks [174]. First, the risk of critical system failures if, for example, an LLM performs poorly for individuals, unlike those represented in the training data [175]. Second, the risk of adversarial attacks [176, 177]. Therefore, researchers and developers have created tools and methods to assess LLMs' robustness, including adversarial methods like red teaming [58], evaluation toolkits like the *Robustness Gym* [178], benchmark datasets like *ANLI* [179], and open-source platforms for model-and-human-in-the-loop testing like *Dynabench* [180]. Particularly relevant for our purposes is *AdvGLUE* [181], which evaluates LLMs' vulnerabilities to adversarial attacks in different domains using a multi-task benchmark. By quantifying robustness, AdvGLUE facilitates comparisons between LLMs and their various affordances and limitations. However, robustness can be operationalised in different ways, e.g., group robustness, which measures a model's performance across different sub-populations [182]. Therefore, model audits should employ multiple tools and methods to assess robustness.

(3) *Information security,* i.e., how difficult it is to extract training data from the LLM. Several LLM-related risks can be understood as 'information hazards' [30],

[22] A wide range of tools and methods to evaluate LLMs already exists. For an overview, see the report *Holistic Evaluation of Language Models* published by researchers at the Center for Research on Foundation Models [35].

including the risk of compromising privacy by leaking personal data. As demonstrated by [165], adversarial agents can perform *training data extraction attacks* to recover personal information like names and social security numbers. However, not all LLMs are equally vulnerable to such attacks. The memorisation of training data can be minimised through differentially private training techniques [183], but their application generally reduces accuracy [184] and increases training time [151]. Promisingly, it is possible to assess the extent to which an LLM has unintentionally memorised rare or unique training data sequences using metrics such as *exposure* [185]. Testing strategies, like exposure, can be employed at the model level, although that requires auditors to have access to the LLM and its training corpus. Still, assessing LLMs' information security during model audits does not address all information hazards because some risk of correctly inferring sensitive information about users can only be audited on an application level.

(4) *Truthfulness,* i.e., to what extent the LLM can distinguish between the real world and possible worlds. Some LLM-related risks stem from their capacity to provide false or misleading information, which creates less well-informed users and potentially erodes public trust in shared information [30]. Statistical methods struggle to distinguish between factually correct versus plausible but factually incorrect information. That problem is exacerbated by the fact that many LLM training practices, like imitating human text on the web or optimising for clicks, are unlikely to create truthful AI [186].[23] However, during model audits, our concern is not developing truthful AI but evaluating truthfulness. Such audits should focus on evaluating overall truthfulness, not the truthfulness of an individual statement. Yet that does not preclude focusing on multiple aspects, e.g., how frequent falsehoods are on average, and how bad worst-case falsehoods are. One benchmark that measures truthfulness is *TruthfulQA* [187], which generates a percentage score using 817 questions spanning 38 application domains, including healthcare and politics. When evaluating an LLM with the help of TruthfulQA, auditors would get a percentage score on how truthful the model is. However, even a strong performance on TruthfulQA does not imply that an LLM will be truthful in a specialised domain. Nevertheless, such benchmarks offer helpful tools for model audits.

These four characteristics pertain to pre-trained LLMs. However, model audits should also review training datasets. It is well-known that training data gaps or biases create models that perform poorly on different datasets [188]. Training LLMs with biased or incomplete data can cause representational and allocational harms [189]. Therefore, a recent European Parliament report [152] discussed mandating third-party audits of AI-training datasets. Technology providers should prepare for such suggestions potentially becoming legal requirements.

Despite these technical and legal considerations, training datasets are often collected with little curation, supervision, or foresight [190]. While curating 'unbiased' datasets may be impossible, disclosing how a dataset was assembled can suggest its potential biases [191]. Model auditors can use existing tools and methods that interrogate biases in LLMs' pre-trained word embeddings, such as the metrics *DisCo* [192], *SEAT* [193] or *CAT* [194]. So-called *data statements* [195] can provide developers and users with the context required to understand specific models' potential biases. *Data representativeness criterion* [196] can determine how representative[24] a training dataset is, and *manual datasets audits* can be supplemented with *automatic analysis* [197]. The *Text Characterisation Toolkit* [198] permits automatic analysis of how dataset properties impact model behaviour. While the availability of such tools is encouraging, it is important to remain realistic about what dataset audits can achieve. Model audits do not aim to ensure that LLMs are ethical in any global sense. Instead, they contribute to better precision in claims about an LLM's capabilities and inform the design of downstream applications.

Model audits require auditors to have privileged access to LLMs and their training datasets. In the typology provided by Koshiyama et al. [10], this corresponds to medium-level access, whereby auditors have access to an LLM equivalent to its developer, meaning they can manipulate model parameters and review learning procedures and task objectives. Such access is required to assess LLMs' capabilities accurately during model audits. However, in contrast to white-box audits, the access model auditors enjoy is limited to the technical system and does not extend to technology providers' organisational processes.

Some of the characteristics tested for during model audits correspond directly to the social and ethical risks LLMs pose. For example, model audits entail evaluating LLMs according to characteristics like information security and truthfulness, which correspond to information hazards and misinformation hazards, respectively, in Weidinger et al.'s taxonomy [30]. Yet it should be noted that our proposed model audits only focus on a few characteristics of LLMs. That is because the criterion of meaningful

[23] Alternative techniques that are better suited for developing truthful AI include bootstrapping, adversarial training [295] and transparent AI [296].

[24] The term 'representativeness' has different meaning in statistics, politics, and machine learning, ranging from a proportionate match between sample and population to a more general sense of inclusiveness [297].

operationalisability sets a high bar: not all risks associated with LLMs can be addressed at the model level. Consider discrimination as an example. Model audits can expose the root causes of some discriminatory practices, such as biases in training datasets that reflect historic injustices. However, what constitutes unjust discrimination is context-dependent and varies between jurisdictions. That problematises saying anything meaningful about risks like unjust discrimination on a model level [199]. While important, that observation does not argue against model audits but for complementary approaches like application audits, as discussed next.

4.4 Application audits

Products and services built using LLMs should undergo application audits that assess the legality of their intended functions and how they will impact users and societies. Unlike governance and model audits, application audits focus on actors employing LLMs in downstream applications. Such audits are well-suited to ensure compliance with national and regional legislation, sector-specific standards, and organisational ethics principles.

Application audits have two components: *functionality audits*, which evaluate applications using LLMs based on their intended and operational goals, and *impact audits*, which evaluate applications based on their impacts on different users, groups, and the natural environment. As discussed in Sect. 3.2, both functionality and impact audits are well-established practices [200]. Next, we consider how they can be combined into procedures for auditing applications based on LLMs.

During *functionality audits*, auditors should check whether the intended purpose of a specific application is (1) legal and ethical in and of itself and (2) aligned with the intended use of the LLM in question. The first check is for legal and ethical compliance, i.e., the adherence to the laws, regulations, guidelines, and specifications relevant to a specific application [201], as well as to voluntary ethics principles [202] or codes of conduct [203]. The purpose of these compliance checks is straightforward: if an application is unlawful or unethical, the performance of its LLM component is irrelevant, and the application should not be permitted on the market.

The second check within functionality audits aim to address the risks stemming from developers overstating or misrepresenting a specific application's capabilities [204]. To do so, functionality audits build on—and accounts for outputs from—audits on other levels. During governance audits, technology providers are obliged to define the intended and disallowed use cases of their LLMs. During model audits, the limitations of LLMs are documented to inform their adaptation downstream. Using such information, functionality audits should ensure that downstream applications are aligned with a given LLM's intended use cases in ways that take account of the model's limitations. Functionality audits thus combines the elements of compliance and risks audit needed to provide assurance for LLMs (Claim 1).

During *impact audits*, auditors disregard an application's intended purpose and technological design to focus only on how its outputs impact different user groups and the environment. The idea behind impact audits is simple: every system can be understood in terms of its inputs and outputs [142]. However, despite that simplicity, implementing impact audits is notoriously hard. AI systems and their environments co-evolve in non-linear ways [137]. Therefore, the link between an LLM-based application's intended purpose and its actual impact may be neither intuitive nor consistent over time. Moreover, it is difficult to track impacts stemming from indirect causal chains [205, 206]. Consequently, establishing which direct and indirect impacts are considered legally and socially relevant remains a context-dependent question which must be resolved on a case-by-case basis. The application must be redesigned or terminated if the impact is considered unacceptable.

Importantly, impact audits should include both pre-deployment (ex-ante) assessments and post-deployment (ex-post) monitoring (Claim 7).[25] The former leverages either empirical evidence or plausible scenarios, depending on how well-defined the application is and the predictability of the environments in which it will operate. For example, applications can be tested in *sandbox environments* [207] that mimic real-world environments and allow developers and policymakers to understand the potential impact before an application goes to market. When used for ML-based systems, sandboxes have proven safe harbours in which to detect and mitigate biases [208]. However, real-world environments often differ from training and testing environments in unforeseen ways [209]. Hence, pre-deployment assessments of LLM-based applications must also use analytical strategies to anticipate the application's impact, e.g., *ethical impact assessments* [110, 210, 211] and *ethical foresight analysis* [153].

Pre-deployment impact assessments and post-deployment monitoring are both individually necessary. As policymakers are well-aware, capturing the full range of potential harms from LLM-based applications requires auditing procedures to include elements of continuous oversight (again, see Claim 7). For example, the EU AI Act requires technology providers to document and analyse high-risk AI systems' performance throughout their life cycles [43]. Methodologically, post-deployment monitoring can be done in

[25] This structure mirrors the 'conformity assessments' and 'post-market monitoring plans' proposed in the EU AI Act [83].

different ways, e.g., periodically reviewing the output from an application and comparing it to relevant standards. Such procedures can also be automated, e.g., by using oversight programs [212] that continuously monitor and evaluate system outputs and alert or intervene if they transgress predefined tolerance spans. Such monitoring can be done by both private companies and government agencies [213]. Overall, application audits seek to ensure that ex-ante testing and impact assessments have been conducted following existing best practices; that post-market plans have been established to enable continuous monitoring of system outputs; and that procedures are in place to mitigate or report different types of failure modes.

By focusing on individual use cases, application audits are well-suited to alerting stakeholders to risks that require much contextual information to understand and address. This includes risks related to discrimination and human–computer interaction harms in Weidinger et al.'s taxonomy [30]. Application audits help identify and manage such risks in several ways. For example, quantitative assessments linking prompts with outputs can give a sense of what kinds of language an LLM is propagating and how appropriate that communication style and content is in different settings [214, 215]. Moreover, qualitative assessments (e.g., those based on interviews and ethnographic methods) can provide insights into users' lived experiences of interacting with an LLM [73].

However, despite those methodological affordances, it remains difficult to define some forms of harm in any global sense [216]. For example, several studies have documented situations in which LLMs propagate toxic language [150, 217], but the interpretation of toxicity and the materialisation of its harms vary across cultural, social, or political groups [218-220]. Sometimes, 'detoxifying' an LLM may be incompatible with other goals and potentially suppress texts written about or by marginalised groups [221]. Moreover, certain expressions might be acceptable in one setting but not in another. In such circumstances, the most promising way forward is to audit not LLMs themselves but downstream applications—thereby ensuring that each application's outputs adhere to contextually appropriate conversational conventions [101].

Another example concerns harmfulness, i.e., the extent to which an LLM-based application inflicts representational, allocational or experiential harms.[26] An LLM that lacks robustness or performs poorly for some social groups may permit unjust discrimination [30] or violate capability fairness [222] when informing real-world allocational decisions like hiring. Multiple benchmarks exist to assess model stereotyping of social groups, including *CrowS-Pairs* [223], *StereoSet* [194] or *Winogender* [224]. To assess risks from experiential harms, quantitative assessments of LLM outputs give a sense of the language it is propagating. For example, [150] have developed the *RealToxicityPrompts* benchmark to assess the toxicity of a generated completion.[27] However, the tools mentioned above are only examples. The main point here is that representational, allocational and experiential harms associated with LLMs are best assessed at the application level through functionality and impact audits as described in this section.

To conduct application audits, lower levels of access are sufficient. For example, to make quantitative assessments to determine the relationship between inputs and outputs, it is sufficient that auditors have what Koshiyama et al. [10] refer to as black-box model access or, in some cases, input data access. Similarly, to audit LLM-based applications for legal compliance and ethical alignment, auditors do not require direct access to the underlying model but can rely on publicly available information—including the claims technology providers and downstream developers make about their systems and the user instructions attached to them.

We contend that governance audits and model audits should be obligatory for all technology providers designing and disseminating LLMs. However, we recommend that application audits should be employed more selectively. Further, although application audits may form the basis for certification [225], auditing does not equal certification. Certification requires predefined standards against which a product or service can be audited and institutional arrangements to ensure the certification process's integrity [131]. Even when not related to certification, application audits' results should be publicly available (at least in summary form). Registries publishing such results incentivise companies to correct behaviour, inform enforcement actions and help cure informational asymmetries in technology regulation [12].

4.5 Connecting the dots

In order to make a real difference to the ways in which LLMs are designed and used, governance, model, and application audits must be connected into a structured process. In practice, this means that outputs from audits on one level become inputs for audits on other levels. Model audits, for instance, produce reports summarising LLMs' properties and limitations, which should inform application audits that verify whether a model's known limitations have been considered when designing downstream applications. Similarly, ex-post application audits produce output logs documenting

[26] [298] distinguish between representational harms (portraying some groups more favourably than others) and allocation harms (allocating resources or opportunities unfairly by social group).

[27] This benchmark relies on *PerspectiveAPI* to score 'toxicity', which is a limitation given that system's weaknesses [145, 299, 300].

the impact that different applications have in applied settings. Such logs should inform LLMs' continuous redesign and revisions of their accompanying model cards. Finally, governance audits must check the extent to which technology providers' software development processes and quality management systems include mechanisms to incorporate feedback from application audits. Figure 2 illustrates how governance, model, and application audits are interconnected in our blueprint.

Each step in our three-layered approach should involve independent third-party auditors (Claim 2). However, two caveats are required here. First, it need not be the same organisation conducting audits on all three levels as each requires different competencies. Governance audits require understanding corporate governance [226] and soft skills like stakeholder communication. Model audits are highly technical and require knowledge about evaluating ML models, operationalising different normative dimensions, and visualising model characteristics. Application auditors typically need domain-specific expertise. All these competencies may not be found within one organisation.

Second, as institutional arrangements vary between jurisdictions and sectors, the best option may be to leverage the capabilities of institutions operating within a specific geography or industry to perform various elements of governance, model, and application audits. For example, medical devices are already subject to various testing and certification procedures before being launched. Hence, application audits for new medical devices incorporating LLMs could be integrated with such procedures. In part, this is already happening. The US Food and Drug Administration (FDA) has proposed a regulatory framework for modifying ML-based software as a medical device [227]. The point is that different independent auditors can perform the three different types of audits outlined here and that different institutional arrangements may be preferable in different jurisdictions or sectors.

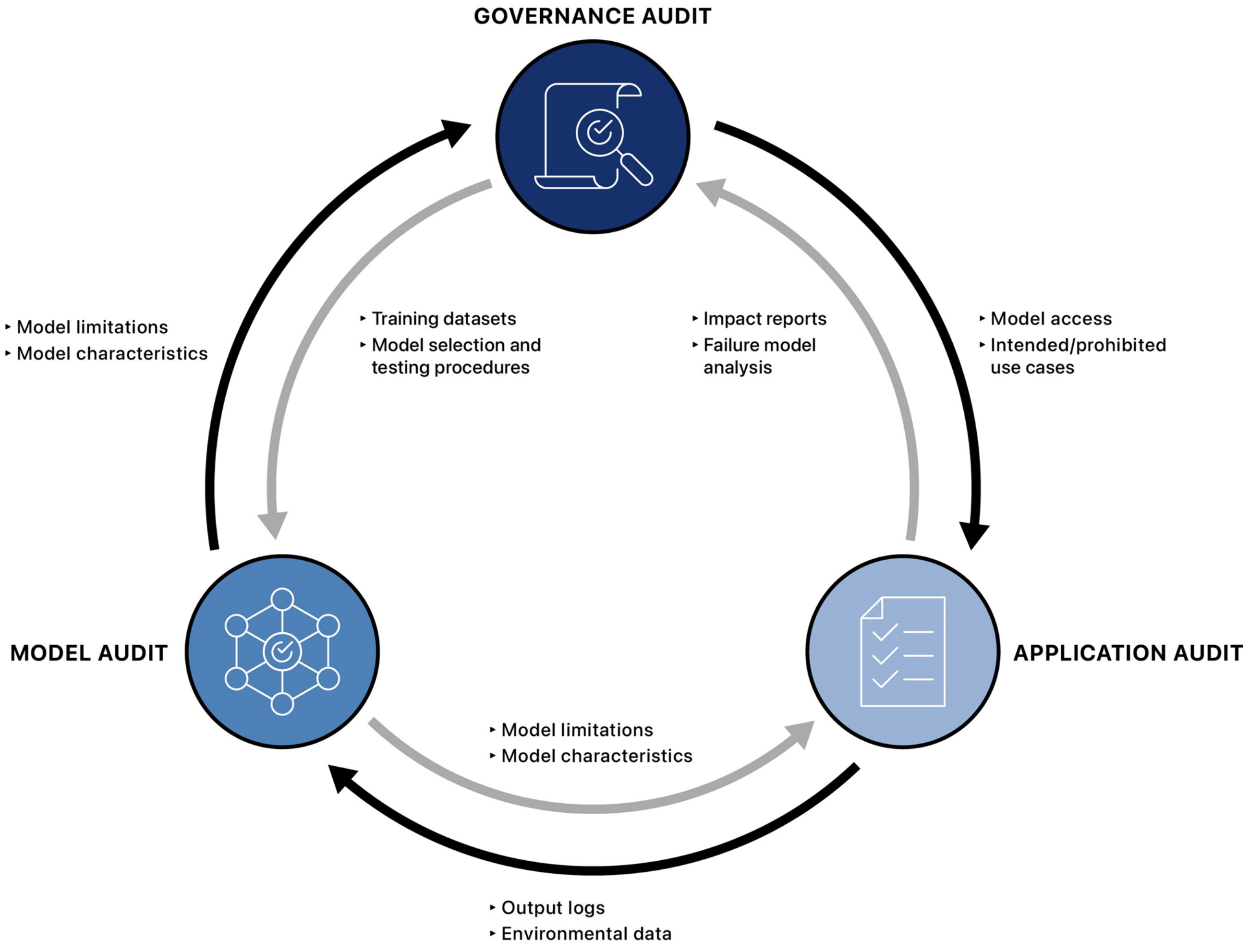


Fig. 2 Outputs from audits on one level become inputs for audits on other levels

5 Limitations and avenues for further research

This section highlights three limitations of our work that apply to any attempt to audit LLMs: one conceptual, one institutional and one practical. First, model audits pose conceptual problems related to construct validity. Second, an institutional ecosystem to support independent third-party audits has yet to emerge. Third, not all LLM-related social and ethical risks can be practically addressed on the technology level. We consider these limitations in turn, discuss potential solutions, and provide directions for future research.

5.1 Lack of methods and metrics to operationalise normative concepts

One bottleneck to developing effective auditing procedures is the difficulty of operationalising normative concepts like robustness and truthfulness [228]. A recent case study found that organisations' lack of standardised evaluation metrics is a crucial challenge when implementing AI auditing procedures [229, 230]. The problem is rooted in construct validity, i.e., the extent to which a given metric accurately measures what it is supposed to [231]. Construct validity problems primarily arise in our blueprint from attempts to operationalise characteristics like performance, robustness, information security and truthfulness during model audits.

Consider truthfulness as an example. LLMs do not require a model of the real world. Instead, they compress vast numbers of conditional probabilities by picking up on language regularities [232, 233]. Therefore, they have no reason to favour any reality but can select from various possible worlds, provided each is internally coherent [234].[28] However, different epistemological positions disagree about the extent to which this way of sensemaking is unique to LLMs or, indeed, a problem at all. Simplifying to the extreme, realists believe in objectivity and the singularity of truth, at least insofar as the natural world is concerned [235]. In contrast, relativists believe that truth and falsity are products of context-dependent conventions and assessment frameworks [236]. Numerous compromise positions can be found on the spectrum between those poles. However, tackling pressing social issues cannot await the resolution of long-standing philosophical disagreements. Indeed, courts settle disagreements daily based on pragmatist operationalisations of concepts like truth and falsehood in keeping with the pragmatic maxim that theories should be judged by their success when applied practically to real-world situations [237].

Following that reasoning, we argue that refining pragmatist operationalisations of concepts like truthfulness and robustness do more to promote fairness, accountability, and transparency in using LLM than either dogmatic or sceptical alternatives [238]. However, developing metrics to capture the essence of thick normative concepts is difficult and entails many well-known pitfalls. Reductionist representations of normative concepts generally bear little resemblance to real-life considerations, which tend to be highly contextual [239]. Moreover, different operationalisations of the same normative concept (like 'fairness') cannot be satisfied simultaneously [240]. Finally, the quantification of normative concepts can itself have subversive or undesired consequences [241, 242]. As Goodhart's Law reminds us, a measure ceases to be a good metric once it becomes a target.

The operationalisation of characteristics like performance, robustness, information security and truthfulness discussed in Sect. 4 is subject to the above limitations. Resolving all construct validity problems may be impossible, but some ways of operationalising normative concepts are better than others for evaluating an LLM's characteristics. Consequently, an important avenue for further research is developing new methods to operationalise normative concepts in ways that are verifiable and maintain high construct validity.

5.2 Lack of an institutional ecosystem

A further limitation is that our blueprint does not decisively identify who should conduct the audits it recommends. This is a limitation, since any auditing procedure will only be as good as the institution delivering it [243]. However, we have left the question open for two reasons. First, different institutional ecosystems intended to support audits and conformity assessments of AI systems are currently emerging in different jurisdictions and sectors [244]. Second, our blueprint is flexible enough to be adopted by any external auditor. Hence, the feasibility and effectiveness of our approach do not hinge on the question of institutional design.

That said, the question of who audits whom is important, and much can be learned from auditing in other domains. Five institutional arrangements for structuring independent audits are particularly relevant to our purposes. Audits of LLMs can be conducted by:

(1) *Private service providers*, chosen by and paid for by the technology provider (equivalent to the role accounting firms play during financial audits or business ethics audits [245]).

(2) *A government agency,* centrally administered and paid for by government, industry, or a combination of both

[28] LLMs favour the statistically most likely reality given their training data. Yet any training data necessarily constitute a reduction of reality that supports some interpretations but obscures others [301].

(equivalent to the FDA's role in approving food and drug substances [246]).[29]

(3) *An industry body*, operationally independent yet funded through fees from its member companies (equivalent to the British Safety Council's role in audits of workers' health and safety [247]).

(4) *Non-profit organisations*, operationally independent and funded through public grants and voluntary donations (equivalent to the Rainforest Alliance role in auditing forestry practices [248]).

(5) *An international organisation*, administered and funded by its member countries (equivalent to the International Atomic Energy Agency's role in auditing nuclear medicine practices [249]).

Each of these arrangements has its own set of affordances and constraints. Private service providers, for example, are under constant pressure to innovate, which can be beneficial given the fast-moving nature of LLM research. However, private providers' reliance on good relationships with technology providers to remain in business increases the risk of collusion [250]. Therefore, some researchers have called for more government involvement, including an 'FDA for algorithms' [251]. Establishing a government agency to review and approve high-risk AI systems could ensure the uniformity and independence of pre-market audits but might stifle innovation and cause longer lead times. Moreover, while the FDA enjoys a solid international reputation [252], not all jurisdictions would consider the judgement of an agency with a national or regional mandate legitimate.

The lack of an institutional ecosystem to implement and enforce the LLM auditing blueprint outlined in this article is a limitation. Without clear institutional arrangements, claims that an AI system has been audited are difficult to verify and may exacerbate harms [133]. Further research could usefully investigate the feasibility and effectiveness of different institutional arrangements for conducting and enforcing the three types of audits proposed.

5.3 Not all risks from LLMs can be addressed on the technology level

Our blueprint for auditing LLMs has been designed to contribute to good governance. However, it cannot eliminate the risks associated with LLMs for three reasons. First, most risks cannot be reduced to zero [125]. Hence, the question is not whether residual risks exist but how severe and socially acceptable they are [253]. Second, some risks stem from deliberate misuse, creating an offensive-defensive asymmetry wherein responsible actors constantly need to guard against all possible vulnerabilities while malicious agents can cause harm by exploiting a single vulnerability [254]. Third, as we will expand on below, not all risks associated with LLMs can be addressed on the technology level.

Weidinger et al. [30] list over 20 risks associated with LLMs divided into six broad risk areas. In Sect. 4, we highlighted how our three-layered approach helps identify and mitigate some of these risks. To recap, governance audits can help protect against risks associated with malicious use; model audits can help identify and manage information and misinformation hazards; and application audits can help protect against discrimination as well as experiential harms. Of course, these are just examples. Audits at each level contribute, directly or indirectly, to addressing many different risks. However, not all the risks listed by Weidinger et al. are captured by our blueprint. Consider automation harm as an example. Increasing the capabilities of LLMs to complete tasks that would otherwise require human intelligence threatens to undermine creative economies [255]. While some highly potent LLMs may remove the basis for some professions that employ many people today—such as translators or copywriters—that is not a failure on the part of the technology. The alternative of building less capable LLMs is counterproductive since abstaining from technology usage generates significant social and economic opportunity costs [256].

The problem is not necessarily change per se but its speed and how the fruits of automation are distributed [257, 258]. Hence, problems related to changing economic environments may be better addressed through social and political reform rather than audits of specific technologies. It is important to remain realistic about auditing's capabilities and not fall into the trap of overpromising when introducing new governance mechanisms [259]. However, the fact that no auditing procedures can address all risks associated with LLMs does not diminish their merits. Instead, it points towards another important avenue for further research: how can and should social and political reform complement technically oriented mechanisms in holistic efforts to govern LLMs?

6 Conclusion

Some of the features that make LLMs attractive also create significant governance challenges. For instance, the potential to adapt LLMs to a wide range of downstream applications undermines system verification procedures that presuppose well-defined demand specifications and predictable operating environments. Consequently, our analysis in Sect. 3 concluded that existing AI auditing procedures are not

[29] Alternative models of government involvement exist. For example, audits may be conducted or sanctioned by a government agency like the National Institute of Standards and Technology (NIST) in the US or by the same notified bodies that the European Commission [43] has tasked with performing conformity assessments of high-risk AI systems in the EU.

well-equipped to assess whether the checks and balances put in place by technology providers and downstream developers are sufficient to ensure good governance of LLMs.

In this article, we have attempted to bridge that gap by outlining a blueprint for how to audit LLMs. In Sect. 4, we introduced a three-layered approach, whereby governance, model and application audits inform and complement each other. During *governance audits*, technology providers' accountability structures and quality management systems are evaluated for robustness, completeness, and adequacy. During *model audits*, LLMs' capabilities and limitations are assessed along several dimensions, including performance, robustness, information security, and truthfulness. Finally, during *application audits*, products and services built on top of LLMs are first assessed for legal compliance and subsequently evaluated based on their impact on users, groups, and the natural environment.

Technology providers and policymakers have already started experimenting with some of the auditing activities we propose. Consequently, auditors can leverage a wide range of existing tools and methods, such as impact assessments, benchmarking, model evaluation, and red teaming, to conduct governance, model, and application audits. That said, the feasibility and effectiveness of our three-layered approach hinge on two factors. First, only when conducted in a combined and coordinated fashion can governance, model and application audits enable different stakeholders to manage LLM-related risks. Hence, audits on the three levels must be connected in a structured process. Governance audits should ensure that providers have mechanisms to take the output logs generated during application audits into account when redesigning LLMs. Similarly, application audits should ensure that downstream developers take the limitations identified during model audits into account when building on top of a specific LLM. Second, audits at each level must be conducted by an independent third-party to ensure that LLMs are ethical, legal, and technically robust. The case for independent audits rests not only on concerns about the misaligned incentives that technology providers may face but also on concerns about the rapidly increasing capabilities of LLMs [260].

However, even when implemented under ideal circumstances, audits will not solve all tensions or protect against all risks of harm associated with LLMs. So, it is important to remain realistic about what auditing can achieve and the main limitations of our approach discussed in Sect. 5 are worth reiterating. To begin with, the feasibility of model audits hinges on the construct validity of the metrics used to assess characteristics like robustness and truthfulness. This is a limitation because such normative concepts are notoriously difficult to operationalise. Further, our blueprint for how to audit LLMs does not specify who should conduct the audits it posits. No auditing procedure is stronger than the institutions backing it. Hence, the fact that an ecosystem of actors capable of implementing our blueprint has yet to emerge constrains its effectiveness. Finally, not all risks associated with LLMs arise from processes that can be addressed through auditing. Some tensions are inherently political and require continuous management through public deliberation and structural reform.

Academics and industry researchers can contribute to overcoming these limitations by focusing on two avenues for further research. The first is to develop new methods and metrics to operationalise normative concepts in ways that are verifiable and maintain a high degree of construct validity. The second is to disentangle further the sources of different types of risks associated with LLMs. Such research would advance our understanding of how political reform can complement technically oriented mechanisms in holistic efforts to govern LLMs.

Policymakers can facilitate the emergence of an institutional ecosystem capable of carrying out and enforcing governance, model, and application audits of LLMs. For example, policymakers can encourage and strengthen private sector auditing initiatives by creating standardised evaluation metrics [261], harmonising AI regulation [262], facilitating knowledge sharing [263] or rewarding achievements through monetary incentives [256]. Policymakers should also update existing and proposed AI regulations in line with our three-layered approach to address LLM-related risks. For example, while the EU AI Act's conformity assessments and post-market monitoring plans mirror application audits, the proposed regulation does not contain mechanisms akin to governance and model audits [83]. Without amendments, such regulations are unlikely to generate adequate safeguards against the risks associated with LLMs.

Our findings most directly concern technology providers as they are primarily responsible for ensuring that LLMs are legal, ethical, and technically robust. Such providers have both moral and material reasons to subject themselves to independent audits, including the need to manage financial and legal risks [264] and build an attractive brand [265]. So, what ought technology providers do? To start with, they should subject themselves to governance audits and their LLMs to model audits. That would create a demand for independent auditing and accreditation bodies and help spark methodological innovation in governance and model audits. Mid-term, Technology providers should also demand that products and services built on top of their LLMs undergo application audits. That could be done through structured access procedures, whereby permission for using an LLM is conditional on such terms. In the long-term, like-minded technology providers should establish, and fund, an independent industry body that conducts or commissions governance, model, and application audits.

Taking a long-term perspective, our three-layered approach holds lessons for how to audit more capable and general future AI systems. This article has focused on LLMs because they have broad societal impacts via widespread applications already today. However, elements of the governance challenges—including generativity, emergence, lack of grounding, and lack of access—have some general applicability to other ML-based systems [266, 267]. Hence, we anticipate that our blueprint can inform the design procedures for auditing other generative, ML-based technologies.

That said, the long-term feasibility and effectiveness of our blueprint for how to audit LLMs may also be undermined by future developments. For example, governance audits make sense when only a limited number of actors have the ability and resources to train and disseminate LLMs. The democratisation of AI capabilities—either through the reduction of entry barriers or a turn to business models based on open-source software—would challenge this status quo [268]. Similarly, if language models become more fragmented or personalised [93], there will be many user-specific branches or instantiations of a single LLM which would make model audits more complex to standardise. As a result, while maintaining the usefulness of our three-layered approach, we acknowledge that it will need to be continuously revised in response to the changing technological and regulatory landscape.

It is worth concluding with some words of caution. Our blueprint is not intended to replace existing governance mechanisms but to complement and interlink them by strengthening procedural transparency and regularity. Rather than being adopted wholesale by technology providers and policymakers, we hope that our three-layered approach can be adopted, adjusted, and expanded to meet the governance needs of different stakeholders and contexts.

Appendix 1: methodology

Before describing our methodology, something should be said about our research approach. According to the pragmatist tradition, research is only legitimate when applied, i.e., grounded in real-world problems [269]. As established in Sect. 2, there is a need to develop new governance mechanisms that different stakeholders can use to identify and mitigate the risks associated with LLMs. In this article, we take a pragmatist stance when exploring how auditing procedures can be designed so they are feasible and effective in practice.

Designing procedures to audit LLMs is an art, not a science. In a policy context, applied research concerns the evaluation of different governance design decisions or policies in relation to a desired outcome [270]. From a pragmatist point of view, however, a mark of quality in applied policy research is that questions are answered in ways that are actionable [237]. That implies that researchers must sometimes go beyond an evaluation of existing options to prescribe new solutions. While there is no guarantee that the best course of action will be found, researchers can ensure rigour by systematically building on previous research and by incorporating input from different stakeholders.

Mindful of those considerations, the following methodology was used to develop our blueprint for how to audit LLMs. Note that while the five steps below exhaust the range of research activities that went into this study, the sequential presentation is a gross simplification. In reality, the research process was messy and iterative, with several of the steps overlapping both thematically and chronologically.

Firstly, we mapped existing auditing procedures designed to identify the risks associated with different AI systems through a systematised literature review [271]. In doing so, we searched five databases (Google Scholar, Scopus, SSRN, Web of Science and arXiv) for articles related to the auditing of AI systems. Keywords for the search included ("auditing, "evaluation" OR "assessment") AND ("fairness", "truthfulness", "transparency" OR "robustness") AND ("language models", "artificial intelligence" OR "algorithms"). However, not all relevant auditing procedures have been developed by academic researchers. Hence, we used a snowballing technique [272], i.e., tracking the citations of already included articles, to identify auditing procedures developed by private service providers, national or regional policymakers and industry associations. A total of 126 documents were included in this systematised literature review.

Secondly, we identified the elements underpinning these procedures. This resulted in a typology that distinguishes between different types of audits, e.g., risk and compliance audits; internal and external audits; ex-ante and ex-post audits; as well as between functionality, code, and impact audits. The space of possible auditing procedures consists of all unique combinations between these different elements.

Thirdly, we generated a list of key claims about how auditing procedures for LLMs should be designed so that they are feasible and effective in practice. To do this, we conducted a gap analysis between the governance challenges posed by LLMs on the one hand and the theoretical affordances of existing AI auditing procedures on the other. Our analysis resulted in seven key claims about how auditing procedures should be designed in order to capture the full range of risks posed by LLMs. Those claims are presented and discussed in Sect. 3.

Fourthly, we created a draft blueprint for how to audit LLMs by identifying the smallest set of auditing procedures that satisfied our seven key claims. In practice, not all auditing procedures are equally effective in identifying the risks posed by LLMs. Besides, some auditing procedures serve similar functions. Although some redundancy is an important feature in safety engineering, too much overlap between different auditing regimes can be counterproductive in so far as roles and responsibilities become less clear and scarce resources are being consumed that could otherwise have been more effectively invested elsewhere. This step thus consisted of reducing the theoretical space of possible auditing procedures into a limited set of activities that are (1) jointly sufficient to identify the full range of risks associated with LLMs, (2) practically feasible to implement, and (3) seem to have a justifiable cost–benefit ratio.

Fifthly and finally, we sought to refine and validate our draft blueprint by triangulating findings [273] from different sources. For example, we sought input from a diverse set of stakeholders. In total, we conducted over 20 semi-structured interviews [274] with, and received feedback from, researchers, professional auditors, AI developers at frontier labs and policymakers in different jurisdictions. The final blueprint outlined in Sect. 4 is the result of those consultations.

Acknowledgments The authors would like to thank the following people for their helpful comments on earlier versions of this manuscript: Markus Anderljung, Matthew Salganik, Arvind Narayanan, Deep Ganguli, Katherine Lee, Toby Shevlane, Varun Rao, and Lennart Heim. In the process of writing the article, the authors also benefited from conversations with and input from Allan Dafoe, Ben Garfinkel, Anders Sandberg, Emma Bluemke, Andreas Hauschke, Owen Larter, Sebastien Krier, Mihir Kshirsagar, and Jade Leung. The article is much better for it.

Author contributions JM is the first author of this paper. JS, HRK, and LF contributed equally to the paper.

Funding JM's doctoral research at the Oxford Internet Institute is supported through a studentship provided by AstraZeneca. JM also conducted part of this research during a paid Research Fellowship at the Centre for Governance of AI.

Data availability We confirm that the research presented in this article is our own, that it has not been submitted to any other journal, and that all sources we used have been credited both in the text and in the bibliography. As part of this research, we have not collected or stored any personal or sensitive data.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

1. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms. In: ICA 2014 Data and Discrimination Preconference, pp. 1–23 (2014). https://doi.org/10.1109/DEXA.2009.55
2. Diakopoulos, N.: Algorithmic accountability: journalistic investigation of computational power structures. Digit. J. **3**(3), 398–415 (2015). https://doi.org/10.1080/21670811.2014.976411
3. Mökander, J., Floridi, L.: Ethics—based auditing to develop trustworthy AI. Minds Mach. (Dordr) **0123456789**, 2–6 (2021). https://doi.org/10.1007/s11023-021-09557-8
4. Brundage, M., et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims. ArXiv, no. 2004.07213[cs.CY])., 2020, [Online]. http://arxiv.org/abs/2004.07213
5. Raji, I.D., Buolamwini, J.: Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: AIES 2019—Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435, (2019). https://doi.org/10.1145/3306618.3314244
6. Mökander, J., Morley, J., Taddeo, M., Floridi, L.: Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. Sci. Eng. Ethics (2021). https://doi.org/10.1007/s11948-021-00319-4ORIGINAL
7. Cobbe, J., Lee, M.S.A., Singh, J.: Reviewable automated decision-making: a framework for accountable algorithmic systems. In: FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 598–609 (2021). https://doi.org/10.1145/3442188.3445921
8. Floridi, L.: Infraethics–on the conditions of possibility of morality. Philos. Technol. **30**(4), 391–394 (2017). https://doi.org/10.1007/s13347-017-0291-1
9. Raji, I.D. et al.: Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: FAT* 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44, 2020, doi: https://doi.org/10.1145/3351095.3372873
10. Koshiyama, A., Kazim, E., Treleaven, P.: Algorithm auditing: managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. IEEE **55**(4), 40–50 (2022). https://doi.org/10.1109/MC.2021.3067225
11. Power, M.: The Audit Society: Rituals of Verification. Oxford University Press, Oxford (1997)
12. Raji, I.D., Xu, P., Honigsberg, C., Ho, D.: Outsider oversight: designing a third party audit ecosystem for AI governance. In: AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557–571, Jul. 2022, https://doi.org/10.1145/3514094.3534181
13. Kazim, E., Koshiyama, A.S., Hilliard, A., Polle, R.: Systematizing audit in algorithmic recruitment. J. Intell. **9**(3), 1–11 (2021). https://doi.org/10.3390/jintelligence9030046
14. Robertson, R.E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., Wilson, C.: Auditing partisan audience bias within Google search. In: Proc ACM Hum Comput Interact, vol. 2, no. CSCW, 2018, https://doi.org/10.1145/3274417
15. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability, and Transparency, 2018, pp. 1–15. https://doi.org/10.2147/OTT.S126905
16. Oakden-Rayner, L., et al.: Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit. Health **4**(5), e351–e358 (2022). https://doi.org/10.1016/S2589-7500(22)00004-8

17. Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K., Oakden-Rayner, L.: The medical algorithmic audit. Lancet Digit. Health **4**(5), e384–e397 (2022). https://doi.org/10.1016/S2589-7500(22)00003-6
18. Bommasani, R., et al.: On the opportunities and risks of foundation models. ArXiv, Aug. 2021, [Online]. http://arxiv.org/abs/2108.07258
19. Bommasani, R., Liang, P.: Reflections on foundation models. HAI, 2021. https://hai.stanford.edu/news/reflections-foundation-models. Accessed 13 Feb 2023
20. Floridi, L., Chiriatti, M.: GPT-3: its nature, scope, limits, and consequences. Minds Mach. **30**(4), 681–694 (2020). https://doi.org/10.1007/s11023-020-09548-1
21. Rosenfeld, R.: Two decdes of statistical language modeling where do we go form here? Where do we go from here? Proc. IEEE **88**(8), 1270–1275 (2000). https://doi.org/10.1109/5.880083
22. Brown, T.B., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. (2020). https://doi.org/10.48550/arxiv.2005.14165
23. OpenAI.: GPT-4 Technical Report. Mar. 2023. [Online]. https://arxiv.org/abs/2303.08774v3. Accessed 12 Apr 2023
24. Chowdhery, A., et al.: PaLM: scaling language modeling with pathways. ArXiv (2022). https://doi.org/10.48550/arxiv.2204.02311
25. Thoppilan, R., et al.: LaMDA: language models for dialog applications. ArXiv (2022)
26. Rae, J.W. et al.: Scaling language models: methods, analysis & insights from training Gopher. ArXiv (2022)
27. S. Zhang *et al.*, "OPT: Open Pre-trained Transformer Language Models," *ArXiv*, May 2022, [Online]. Available: http://arxiv.org/abs/2205.01068
28. Hu, Y., Jing, X., Ko, Y., Rayz, J.T.: Misspelling correction with pre-trained contextual language model. In: 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCICC), pp. 144–149. (2020)
29. Hsieh, K.: Transformer poetry: poetry classics reimagined by artificial intelligence. San Francisco: Paper Gains Publishing, 2019. [Online]. https://papergains.co/pdfs/Transformer_Poetry-978-1-7341647-0-1.pdf. Accessed 20 Jan 2023
30. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, A.L., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., Gabriel, I.: Taxonomy of risks posed by language models. In: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, pp. 214–229. New York, NY, USA. (2022). https://doi.org/10.1145/3531146.3533088
31. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Inc, Mar. 2021, pp. 610–623. https://doi.org/10.1145/3442188.3445922
32. Shelby, R., et al.: Sociotechnical harms: scoping a taxonomy for harm reduction. ArXiv (2022). https://doi.org/10.48550/arxiv.2210.05791
33. Perez, E., et al.: Discovering language model behaviors with model-written evaluations. ArXiv, Dec. 2022. [Online]. https://arxiv.org/abs/2212.09251v1. Accessed 22 Mar 2023
34. Curry, D.: ChatGPT Revenue and Usage Statistics (2023)—Business of Apps. *BusinessofApps*, 2023. [Online]. https://www.businessofapps.com/data/chatgpt-statistics/. Accessed 2 Apr 2023
35. Liang, P., et al.: Holistic evaluation of language models; holistic evaluation of language models. ArXiv, 2022. [Online]. https://arxiv.org/pdf/2211.09110.pdf. Accessed 13 Feb 2023
36. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 2021. [Online]. https://github.com/OpenAI/CLIP. Accessed 20 Jan 2023
37. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M, OpenAI.: Hierarchical text-conditional image generation with CLIP latents. 2022. https://doi.org/10.48550/arxiv.2204.06125
38. OpenAI.: Best practices for deploying language models. Website, 2022. https://openai.com/blog/best-practices-for-deploying-language-models/. Accessed 20 Jan 2023
39. Peyrard, M., et al.: Invariant language modeling. ArXiv (2021). https://doi.org/10.48550/arxiv.2110.08413
40. Ganguli, D., et al.: Predictability and surprise in large generative models. In: ACM International Conference Proceeding Series, pp. 1747–1764, Jun. 2022. https://doi.org/10.1145/3531146.3533229.
41. Ganguli, D., et al.: Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. ArXiv, 2022. [Online]. https://github.com/anthropics/hh-rlhf. Accessed 2 Apr 2023
42. Perez, E., et al.: Red teaming language models with language models. ArXiv (2022). https://doi.org/10.48550/arxiv.2202.03286
43. European Commission: Proposal for regulation of the European parliament and of the council—Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
44. Office of U.S. Senator Ron Wyden.: Algorithmic Accountability Act of 2022. In: 117th Congress 2D Session, 2022, https://doi.org/10.1016/S0140-6736(02)37657-8
45. Joshi, A.K.: Natural language processing. Science (1979) **253**(5025), 1242–1249 (1991). https://doi.org/10.1126/SCIENCE.253.5025.1242
46. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science (1979) **349**(6245), 261–266 (2015). https://doi.org/10.1126/SCIENCE.AAA8685/ASSET/D33AB763-A443-444C-B766-A6B69883BFD7/ASSETS/GRAPHIC/349_261_F5.JPEG
47. Chernyavskiy, A., Ilvovsky, D., Nakov, P.: Transformers: 'The End of History' for Natural Language Processing?," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12977 LNAI, pp. 677–693, 2021. https://doi.org/10.1007/978-3-030-86523-8_41/TABLES/5
48. Adiwardana, D., et al.: Towards a human-like open-domain Chatbot. ArXiv (2020). https://doi.org/10.48550/arxiv.2001.09977
49. Wang, A., et al.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems.In: NIPS'19, 2019. https://doi.org/10.5555/3454287.3454581
50. Bai, Y., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. ArXiv, Apr. 2022. [Online]. https://arxiv.org/abs/2204.05862v1. Accessed 2 Apr 2023
51. Stiennon, N., et al. Learning to summarize from human feedback. Adv Neural Inf Process Syst. vol. 2020-December, Sep. 2020. [Online]. https://arxiv.org/abs/2009.01325v3. Accessed 2 Apr 2023
52. Ouyang, L., et al.: Training language models to follow instructions with human feedback. ArXiv, Mar. 2022. [Online]. https://arxiv.org/abs/2203.02155v1. Accessed 2 Apr 2023
53. Arcas, B.A.Y.: Do large language models understand us? Daedalus **151**(2), 183–197 (2022). https://doi.org/10.1162/daed_a_01909

54. Suzgun, M., et al.: Challenging BIG-bench tasks and whether chain-of-thought can solve them. ArXiv, Oct. 2022. [Online]. https://arxiv.org/abs/2210.09261v1. Accessed 2 Apr 2023
55. Villalobos, P., Sevilla, J., Besiroglu, T., Heim, L., Ho, A., Hobbhahn, M.: Machine learning model sizes and the parameter gap. ArXiv, Jul. 2022, [Online]. http://arxiv.org/abs/2207.02852
56. Hoffmann, J. et al.: Training compute-optimal large language models. ArXiv, Mar. 2022, [Online]. http://arxiv.org/abs/2203.15556
57. Bowman, S.R.: Eight things to know about large language models. Apr. 2023. [Online]. https://arxiv.org/abs/2304.00612v1 Accessed 12 Apr 2023
58. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. **4**(1), 1–13 (2021). https://doi.org/10.1038/s41746-021-00455-y
59. Wang, Y., Wang, W., Joty, S., Hoi, S.C.H.: CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In: EMNLP 2021—2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, pp. 8696–8708, 2021. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.685
60. Chen, M., et al.: Evaluating large language models trained on code. ArXiv. [Online]. https://www.github.com/openai/human-eval (2021). Accessed 20 Jan 2023
61. Wang, S., Tu, Z., Tan, Z., Wang, W., Sun, M., Liu, Y.: Language models are good translators. ArXiv (2021). https://doi.org/10.48550/arxiv.2106.13627
62. Kojima, T., Shane Gu, S., Reid, M., Matsuo, Y., Iwasawa, Y., Google Research: Large language models are zero-shot reasoners. ArXiv (2022). https://doi.org/10.48550/arxiv.2205.11916
63. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020). https://doi.org/10.48550/arXiv.2001.08361
64. Brown, A.R., Kluska, A., Nie, A., Gupta, A., Venkatesh, A., Gholamidavoodi, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (2022). arXiv:2206.04615
65. Kirk, H.R., et al.: Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. NeurIPS, 2021. [Online]. https://github.com/oxai/intersectional_gpt2. Accessed 20 Jan 2023
66. Azaria, A.: ChatGPT Usage and Limitations. HAL, Dec. 2022. [Online]. https://hal.science/hal-03913837. Accessed 2 Apr 2023
67. Borji, A., Ai, Q.: A categorical archive of ChatGPT failures. Feb. 2023. [Online]. https://arxiv.org/abs/2302.03494v7. Accessed 22 Mar 2023
68. Shevlane, T.: Structured access. In: Bullock, J., Chen, Y.-C., Himmelreich, J., Hudson, V.M., Korinek, A., Young, M., Zhang, B. (eds.) The Oxford Handbook of AI Governance. Oxford University Press (2022). https://doi.org/10.1093/oxfordhb/9780197579329.013.39
69. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models. ArXiv, Feb. 2021, [Online]. http://arxiv.org/abs/2102.02503
70. Avin, S., et al.: Filling gaps in trustworthy development of AI. Science **374**(6573), 1327–1329 (2021). https://doi.org/10.1126/SCIENCE.ABI7176
71. PAI.: Researching Diversity, Equity, and Inclusion in the Field of AI-Partnership on AI. Website, 2020. https://partnershiponai.org/researching-diversity-equity-and-inclusion-in-the-field-of-ai/. Accessed 20 Jan 2023
72. Wang, Z.J., Choi, D., Xu, S., Yang, D.: Putting humans in the natural language processing loop: a survey. In: Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pp. 47–52 (2021)
73. Marda, V., Narayan, S.: On the importance of ethnographic methods in AI research. Nat. Mach. Intell. **3**(3. Nature Research), 187–189 (2021). https://doi.org/10.1038/s42256-021-00323-0
74. Mitchell, M., et al.: Model cards for model reporting. In: FAT* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pp. 220–229, Jan. 2019. https://doi.org/10.1145/3287560.3287596
75. Derczynski, L., et al.: Assessing language model deployment with risk cards. [Online]. https://arxiv.org/abs/2303.18190v1 (2023). Accessed 2 Apr 2023
76. Gebru, T., et al.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021). https://doi.org/10.1145/3458723
77. MetaAI.: System Cards, a new resource for understanding how AI systems work. Website. https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/ (2023). Accessed 20 Jan 2023
78. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models (2023). arXiv:2301.10226
79. Hacker, P., Engel, A., Mauer, M.: Regulating ChatGPT and other large generative AI models (2023). arXiv:2302.02337
80. Engler, A.: Early thoughts on regulating generative AI like ChatGPT. In: Brookings TechTank. https://www.brookings.edu/blog/techtank/2023/02/21/early-thoughts-on-regulating-generative-ai-like-chatgpt/ (2023). Accessed 2 Apr 2023
81. Altman, S.: Planning for AGI and beyond. OpenAI Blog. [Online]. https://openai.com/blog/planning-for-agi-and-beyond#fn1 (2023). Accessed 24 Mar 2023
82. Helberger, N., Diakopoulos, N.: ChatGPT and the AI Act. Internet Policy Rev. (2023). https://doi.org/10.14763/2023.1.1682
83. Mökander, J., Axente, M., Casolari, F., Floridi, L.: Conformity assessments and post-market monitoring: a guide to the role of auditing in the Proposed European AI regulation. Minds Mach. (Dordr) **32**(2), 241–268 (2022). https://doi.org/10.1007/s11023-021-09577-4
84. Lee, T.-H., Azham, M.A.: The evolution of auditing: An analysis of the historical development. [Online]. https://www.researchgate.net/publication/339251518 (2008). Accessed 10 Feb 2023
85. Senft, S., Gallegos, F.: Information Technology Control and Audit, 3rd edn. CRC Press/Auerbach Publications, Boca Raton (2009)
86. Dai, W., Berleant, D.: Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics. In: Proceedings—2019 IEEE 1st International Conference on Cognitive Machine Intelligence, CogMI 2019, pp. 148–155, Dec. 2019. https://doi.org/10.1109/COGMI48466.2019.00029.
87. Voas, J., Miller, K.: Software certification services: encouraging trust and reasonable expectations. In: IEEE Computer Society, pp. 39–44. [Online]. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1717342 (2016). Accessed 2 Apr 2023
88. Dean, S., Gilbert, T.K., Lambert, N., Zick, T.: Axes for Sociotechnical Inquiry in AI Research. IEEE Trans. Technol. Soc. **2**(2), 62–70 (2021). https://doi.org/10.1109/tts.2021.3074097
89. European Commission.: AI liability directive. In: Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence, pp. 1–29. [Online]. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf (2022). Accessed 21 Jan 2023
90. Berlin, I.: The pursuit of the ideal. In: The Crooked Timber of Mankind: Chapters in the History of Ideas, 1988, pp. 1–20. [Online]. https://www-jstor-org.ezproxy-prd.bodleian.ox.ac.uk/stable/j.ctt2tt8nd.6#metadata_info_tab_contents. Accessed 20 Jan 2023

91. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. (Dordr) **30**(3), 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2
92. Goodman, B.: Hard choices and hard limits in artificial intelligence. In: AIES 2021-Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 112–120, Jul. 2021, https://doi.org/10.1145/3461702.3462539
93. Kirk, H.R., Vidgen, B., Röttger, P., Hale, S.A.: Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. ArXiv (2023)
94. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
95. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: a survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) (2021). https://doi.org/10.1145/3457607
96. Kleinberg, J.: Inherent trade-offs in algorithmic fairness. ACM SIGMETRICS Perform. Eval. Rev. **46**(1), 40–40 (2018). https://doi.org/10.1145/3292040.3219634
97. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: 31st Conference on Neural Information Processing Systems. [Online]. https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data (2017). Accessed 20 Jan 2023
98. Whittlestone, J., Alexandrova, A., Nyrup, R., Cave, S.: The role and limits of principles in AI ethics: Towards a focus on tensions. In: AIES 2019—Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 195–200 (2019). doi: https://doi.org/10.1145/3306618.3314289
99. Gururangan, S, et al.: Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360, Jul. 2020, https://doi.org/10.18653/V1/2020.ACL-MAIN.740
100. O'Neill, O.: A Philosopher Looks at Digital Communication. Cambridge University Press, Cambridge (2021)
101. Kasirzadeh, A., Gabriel, I.: In conversation with Artificial Intelligence: aligning language models with human values. Minds Mach. (Dordr) (2023). https://doi.org/10.48550/arxiv.2209.00731
102. Steed, R., Panda, S., Kobren, A., Wick, M.: Upstream mitigation is not all you need: testing the bias transfer hypothesis in pre-trained language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 3524–3542 (2022). https://doi.org/10.18653/V1/2022.ACL-LONG.247
103. Gupta, K.: Comtemporary Auditing," p. 1095. [Online]. https://books.google.com/books/about/Contemporary_Auditing.html?id=neDFWDyUWuQC (2004). Accessed 18 Feb 2023
104. Flint, D.: Philosophy and Principles of Auditing: An Introduction. Macmillan Education, Basingstoke (1988)
105. LaBrie, R.C., Steinke, G.H.: Towards a framework for ethical audits of AI algorithms. In: 25th Americas Conference on Information Systems, AMCIS 2019, pp. 1–5 (2019)
106. Stodt, J., Reich, C.: Machine learning development audit framework: assessment and inspection of risk and quality of data, model and development process. Int. J. Comput. Inform. Eng. **15**(3), 187–193, (2021)
107. Adler, P., et al.: Auditing black-box models for indirect influence. Knowl. Inf. Syst. **54**(1), 95–122 (2018). https://doi.org/10.1007/s10115-017-1116-3
108. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: 35th International Conference on Machine Learning, ICML 2018, vol. 6, pp. 4008–4016 (2018)
109. Laux, J., Wachter, S., Mittelstadt, B.: Taming the few: platform regulation, independent audits, and the risks of capture created by the DMA and DSA. Comput. Law Secur. Rev. **43**, 105613 (2021). https://doi.org/10.1016/j.clsr.2021.105613
110. Selbst, A.D.: An institutional view of algorithmic impact assessments. Harv. J. Law Technol., vol. 35, 2021, [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634. Accessed 10 Feb 2023
111. Bandy, J.: Problematic machine behavior: a systematic literature review of algorithm audits. Proc. ACM Hum. Comput. Interact. **5**(1), 1–34 (2021). https://doi.org/10.1145/3449148
112. Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., Wen, Y.: capAI—a procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. SSRN (2022). https://doi.org/10.2139/ssrn.4064091
113. Minkkinen, M., Laine, J., Mäntymäki, M.: Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks. Digit. Soc. **1**(3), 21 (2022). https://doi.org/10.1007/s44206-022-00022-2
114. Metaxa, D., et al.: Auditing algorithms. Found. Trends Human-Comput. Interact. **14**(4), 272–344 (2021). https://doi.org/10.1561/1100000083
115. Berghout, E., Fijneman, R., Hendriks, L., de Boer, M., Butijn, B.J.: Advanced digital auditing. In: Progress in IS. Cham: Springer Nature, (2023). https://doi.org/10.1007/978-3-031-11089-4
116. Mökander, J., Axente, M.: Ethics-based auditing of automated decision-making systems : intervention points and policy implications. AI Soc (2021). https://doi.org/10.1007/s00146-021-01286-x
117. Brown, S., Davidovic, J., Hasan, A.: The algorithm audit: scoring the algorithms that score us. Big Data Soc **8**(1), 205395172098386 (2021). https://doi.org/10.1177/2053951720983865
118. Gibson Dunn.: New York City Proposes Rules to Clarify Upcoming Artificial Intelligence Law for Employers. https://www.gibsondunn.com/new-york-city-proposes-rules-to-clarify-upcoming-artificial-intelligence-law-for-employers/ (2023). Accessed 2 Apr 2023
119. PwC: PwC ethical AI framework (2020). https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html. Accessed 10 Feb 2023
120. Deloitte.: Deloitte Introduces Trustworthy AI Framework to Guide Organizations in Ethical Application of Technology. Press release. https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-introduces-trustworthy-ai-framework.html (2020). Accessed 18 Sep 2020)
121. KPMG.: KPMG offers ethical AI Assurance using CIO Strategy Council standards. Press release. https://home.kpmg/ca/en/home/media/press-releases/2020/11/kpmg-offers-ethical-ai-assurance-using-ciosc-standards.html (2020). Accessed 10 Nov 2021
122. EY.: Assurance in the age of AI," 2018, [Online]. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-assurance-in-the-age-of-ai.pdf. Accessed 12 Feb 2023
123. NIST: AI Risk Management Framework: Second Draft Notes for Reviewers: Call for comments and contributions. National Institute of Standards and Technology (2022)
124. ISO.: ISO/IEC 23894-Information technology—Artificial intelligence—Guidance on risk management. International Organization for Standardization. https://www.iso.org/standard/77304.html, (2023). Accessed 20 Jan 2023
125. NIST.: Risk management guide for information technology systems recommendations of the National Institute of Standards and Technology. National Institute of Standards and Technology. [Online]. https://www.hhs.gov/sites/default/files/ocr/priva

cy/hipaa/administrative/securityrule/nist800-30.pdf (2002). Accessed 20 Jan 2023
126. VDE.: VDE SPEC 900012 V1.0 (en). Verband Der Elektrotechnik. [Online]. www.vde.com (2022). Accessed 20 Jan 2023
127. ICO.: Guidance on the AI auditing framework: Draft guidance for consultation. Information Commissioner's Office, [Online]. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (2020). Accessed 12 Feb 2023
128. Institute of Internal Auditors: The IIA's Artificial Intelligence Auditing Framework. The Institute of Internal Auditors-Global Perspectives (2018)
129. ISO.: ISO/IEC 38507:2022-Information technology—Governance of IT—Governance implications of the use of artificial intelligence by organizations. International Organization for Standardization. https://www.iso.org/standard/56641.html?browse=tc (2022). Accessed 20 Jan 2023
130. Institute of Internal Auditors. About Internal Audit. The Institute of Internal Auditors. https://www.theiia.org/en/about-us/about-internal-audit/ (2022). Accessed 20 Jan 2023
131. Yanisky-Ravid, S., Hallisey, S.K.: Equality and privacy by design: a new model of artificial data transparency via auditing, certification, and safe harbor regimes. Fordham Urban Law J. **46**(2), 428-486, (2019).
132. Saleiro, P., et al.: Aequitas: a bias and fairness audit toolkit. ArXiv, no. 2018, 2018, [Online]. http://arxiv.org/abs/1811.05577
133. Costanza-Chock, S., Raji, I.D., Buolamwini, j.: Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem; Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In:FAccT'22, vol. 22 (2022). https://doi.org/10.1145/3531146.3533213
134. Slee, T.: The incompatible incentives of private-sector AI. In: The Oxford Handbook of Ethics of AI, Oxford University Press, pp. 106–123 (2020). https://doi.org/10.1093/OXFORDHB/9780190067397.013.6
135. Naudé, W., Dimitri, N.: The race for an artificial general intelligence: implications for public policy. AI Soc. **35**(2), 367–379 (2020). https://doi.org/10.1007/S00146-019-00887-X/METRICS
136. Engler, A.C.: Outside auditors are struggling to hold AI companies accountable. FastCompany. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue (2021). Accessed 20 Jan 2023
137. Lauer, D.: You cannot have AI ethics without ethics. AI and Ethics **0123456789**, 1–5 (2020). https://doi.org/10.1007/s43681-020-00013-4
138. Danks, D., London, A.J.: Regulating autonomous systems: beyond standards. IEEE Intell Syst **32**(1), 88–91 (2017). https://doi.org/10.1109/MIS.2017.1
139. Mahajan, V., Venugopal, V.K., Murugavel, M., Mahajan, H.: The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. Acad. Radiol. **27**(1), 132–135 (2020). https://doi.org/10.1016/j.acra.2019.09.009
140. Zerbino, P., Aloini, D., Dulmin, R., Mininno, V.: Process-mining-enabled audit of information systems: methodology and an application. Expert Syst. Appl. **110**, 80–92 (2018). https://doi.org/10.1016/j.eswa.2018.05.030
141. Mittelstadt, B.: Auditing for transparency in content personalization systems. Int. J. Commun. **10**(June), 4991–5002 (2016)
142. Kroll, J.A.: The fallacy of inscrutability. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. (2018). https://doi.org/10.1098/rsta.2018.0084
143. OECD.: OECD Framework for the Classification of AI systems. Paris. [Online]. https://doi.org/10.1787/cb6d9eca-en (2022) Accessed 11 Apr 2022
144. Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., Song, D.: Fooling vision and language models despite localization and attention mechanism. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4951–4961, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00520
145. Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.B.: HateCheck: functional tests for hate speech detection models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 41–58 (2021). https://doi.org/10.18653/V1/2021.ACL-LONG.4
146. Aspillaga, C., Carvallo, A., Araujo, V.: Stress Test evaluation of transformer-based models in natural language understanding tasks. In: Proceedings of the 12th Conference on Language Resources and Evaluation, pp. 11–16. [Online]. https://github.com/ (2020). Accessed 20 Jan 2023
147. Dignum, V.: Responsible autonomy. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, vol. 1, p. 5, (2017). .24963/ijcai.2017/655
148. Wei, J. et al.: Emergent abilities of large language models. ArXiv (2022)
149. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: The woman worked as a babysitter: on biases in language generation. In: 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3407–3412 (2019). https://doi.org/10.18653/V1/D19-1339.
150. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N.A.: RealToxicityPrompts: evaluating neural toxic degeneration in language models. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3356–3369, Sep. 2020, [Online]. http://arxiv.org/abs/2009.11462
151. Song, C., Shmatikov, V.: Auditing data provenance in text-generation models. In: KDD'19, 2019. https://doi.org/10.1145/3292500.3330885
152. EPRS: Auditing the quality of datasets used in algorithmic decision-making systems. European Parliamentary Research Service (2022). https://doi.org/10.2861/98930
153. Floridi, L., Strait, A.: Ethical foresight analysis: what it is and why it is needed? Minds Mach. (Dordr) **30**(1), 77–97 (2020). https://doi.org/10.1007/s11023-020-09521-y
154. Hodges, C.: Ethics in business practice and regulation. In: Law and Corporate Behaviour : Integrating Theories of Regulation, Enforcement, Compliance and Ethics, pp. 1–21 (2015). https://doi.org/10.5040/9781474201124
155. ISO.: ISO/IEC 38500:2015—Information technology—Governance of IT for the organization. International Organization for Standardization. https://www.iso.org/standard/62816.html (2015). Accessed 20 Jan 2023).
156. Iliescu, F.-M.: Auditing IT Governance. In: Informatica Economica, **14**(1), 93–102. [Online]. https://www.proquest.com/docview/1433236144/fulltextPDF/A2EAFE83CBFA461APQ/1?accountid=13042&forcedol=true (2010). Accessed 20 Jan 2023
157. Falco, G., et al.: Governing AI safety through independent audits. Nat. Mach. Intell. **3**(7), 566–571 (2021). https://doi.org/10.1038/s42256-021-00370-7
158. Leveson, N.: Engineering a safer world: systems thinking applied to safety. In: Engineering systems. Cambridge: MIT Press (2011)
159. Dobbe, R.I.J.: System safety and artificial intelligence. In: The Oxford Handbook of AI Governance, p. C67.S1-C67.S18, Oct. 2022. https://doi.org/10.1093/OXFORDHB/9780197579329.013.67
160. Schuett, J.: Three lines of defense against risks from AI. (2022). https://doi.org/10.48550/arxiv.2212.08364

161. Bauer, J.: The necessity of auditing artificial intelligence. SSRN **577**, 1–16 (2016)
162. Chopra, A.K., Singh, M.P.: Sociotechnical systems and ethics in the large. In: AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 48–53 (2018). https://doi.org/10.1145/3278721.3278740
163. Contractor, D., et al.: Behavioral use licensing for responsible AI. In" 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM, Jun. 2022, pp. 778–788. https://doi.org/10.1145/3531146.3533143.
164. Schuett, J.: Risk management in the artificial intelligence act. Eur. J. Risk Regul. (2023). https://doi.org/10.1017/ERR.2023.1
165. Carlini, N., et al.: Extracting training data from large language models. In: Proceedings of the 30th USENIX Security Symposium, 2021. [Online]. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. Accessed 20 Jan 2023
166. Dwork, C.: Differential privacy. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4052 LNCS, pp. 1–12 (2006). https://doi.org/10.1007/11787006_1/COVER
167. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. **2**(6), 305–311 (2020). https://doi.org/10.1038/s42256-020-0186-1
168. Bharadwaj, K.B.P., Kanagachidambaresan, G.R.: Pattern recognition and machine learning. (2021). https://doi.org/10.1007/978-3-030-57077-4_11
169. Crisan, A., Drouhard, M., Vig, J., Rajani, N.: Interactive model cards: a human-centered approach to model documentation. In: ACM International Conference Proceeding Series, vol. 22, pp. 427–439, Jun. 2022. https://doi.org/10.1145/3531146.3533108
170. Pushkarna, M., Zaldivar, A., Kjartansson, O.: Data cards: purposeful and transparent dataset documentation for responsible AI. In: ACM International Conference Proceeding Series, pp. 1776–1826, Jun. 2022. https://doi.org/10.1145/3531146.3533231
171. Jernite, Y. et al.: Data governance in the age of large-scale data-driven language technology. In 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM, Jun. 2022, pp. 2206–2222. https://doi.org/10.1145/3531146.3534637
172. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis)contents: a survey of dataset development and use in machine learning research. Patterns **2**(11), 100336 (2021). https://doi.org/10.1016/J.PATTER.2021.100336
173. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop, pp. 353–355 (2018). https://doi.org/10.18653/V1/W18-5446.
174. Rudner, T.J., Toner, H.: Key concepts in AI Safety: robustness and adversarial examples. In: CSET Issue Brief (2021)
175. Sohoni, N.S., Dunnmon, J.A., Angus, G., Gu, A., Ré, C.: No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In: 34th Conference on Neural Information Processing Systems, Nov. 2020, [Online]. http://arxiv.org/abs/2011.12945
176. Garg, S., Ramakrishnan, G.: BAE: BERT-based adversarial examples for text classification. In: EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 6174–6181, (2020). https://doi.org/10.18653/V1/2020.EMNLP-MAIN.498
177. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-ATTACK: adversarial attack against BERT using BERT. In: EMNLP 2020—2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 6193–6202, 2020. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.500.
178. Goel, K., Rajani, N.. Vig, J.. Taschdjian, Z., Bansal, M., Ré, C.: Robustness gym: unifying the NLP evaluation landscape. In: NAACL-HLT 2021—2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Demonstrations, pp. 42–55 (2021). https://doi.org/10.18653/V1/2021.NAACL-DEMOS.6.
179. Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., Kiela, D.: Adversarial NLI: A new benchmark for natural language understanding. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4885–4901, Jul. 2020. https://doi.org/10.18653/V1/2020.ACL-MAIN.441
180. Kiela, D., et al.: Dynabench: rethinking benchmarking in NLP. In: NAACL-HLT 2021—2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 4110–4124, (2021). https://doi.org/10.18653/V1/2021.NAACL-MAIN.324
181. Wang, B., et al.: Adversarial GLUE: a multi-task benchmark for robustness evaluation of language models. In: NeurIPS 2021, Nov. 2021. https://doi.org/10.48550/arxiv.2111.02840.
182. Zhang, M., Ré, C.: Contrastive adapters for foundation model group robustness. In: ICML 2022 Workshop on Spurious Correlations, Jul. 2022. https://doi.org/10.48550/arxiv.2207.07180
183. McMahan, H.B., Ramage, D., Talwar, K., Zhang, L.: Learning differentially private recurrent language models. In: ICLR 2018 Conference Blind Submission. Feb. 24, 2018
184. Jayaraman, B., Evans, D.: Evaluating differentially private machine learning in practice. In: Proceedings of the 28th USENIX Security Symposium, Feb. 2019, [Online]. http://arxiv.org/abs/1902.08874
185. Carlini, N., Brain, G., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Security Symposium, 2019, [Online]. https://www.usenix.org/conference/usenixsecurity19/presentation/carlini. Accessed 10 Feb 2023
186. Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., et al.: Truthful AI: developing and governing AI that does not lie (2021). arXiv:2110.06674
187. Lin, S., Openai, J.H., Evans, O.: TruthfulQA: measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 3214–3252, Jun. 2022. https://doi.org/10.18653/V1/2022.ACL-LONG.229
188. Nejadgholi., I., Kiritchenko, S.: On cross-dataset generalization in automatic detection of online abuse. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 173–183 (2020). https://doi.org/10.18653/v1/P17
189. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science (1979) **356**(6334), 183–186 (2017). https://doi.org/10.1126/SCIENCE.AAL4230/SUPPL_FILE/CALISKAN-SM.PDF
190. Jo, E.S., Gebru, T.: Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 306–316, Jan. 2020. doi: https://doi.org/10.1145/3351095.3372829
191. Dodge, J., et al.: Documenting large Webtext corpora: a case study on the colossal clean crawled corpus. In: EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language

Processing, Proceedings, pp. 1286–1305 (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.98

192. Webster, K., et al.: Measuring and reducing gendered correlations in pre-trained models. ArXiv (2020). https://doi.org/10.48550/arxiv.2010.06032
193. May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. In: NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 622–628 (2019). https://doi.org/10.18653/V1/N19-1063
194. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 5356–5371. [Online]. https://stereoset (2021). Accessed 20 Jan 2023
195. Bender, E.M., Friedman, B.: Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans. Assoc. Comput. Linguist. **6**, 587–604 (2018). https://doi.org/10.1162/TACL_A_00041
196. Schat, E., van de Schoot, R., Kouw, W.M., Veen, D., Mendrik, A.M.: The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. PLoS ONE **15**(8), e0237009 (2020). https://doi.org/10.1371/JOURNAL.PONE.0237009
197. Kreutzer, J., et al.: Quality at a glance: an audit of web-crawled multilingual datasets. Trans. Assoc. Comput. Linguist. **10**, 50–72 (2022). https://doi.org/10.1162/TACL_A_00447/109285
198. Simig, D., et al.: Text Characterization Toolkit (TCT). In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 72–87. [Online]. https://aclanthology.org/2022.aacl-demo.9 (2022). Accessed 21 Jan 2023
199. Hancox-Li, L., Kumar, I.E.: Epistemic values in feature importance methods: Lessons from feminist epistemology. In: FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 817–826, Mar. 2021. https://doi.org/10.1145/3442188.3445943.
200. Dash, A., Mukherjee, A., Ghosh, S.: A network-centric framework for auditing recommendation systems. In: Proceedings—IEEE INFOCOM, vol. April, pp. 1990–1998 (2019). https://doi.org/10.1109/INFOCOM.2019.8737486.
201. Idowu, S.O.: Legal compliance. In: Idowu, S.O., Capaldi, N., Zu, L., Gupta, A.D. (eds.) Encyclopedia of Corporate Social Responsibility, pp. 1578–1578. Springer. Berlin (2013). https://doi.org/10.1007/978-3-642-28036-8_100980
202. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
203. Green, R.M., Donovan, A.: The methods of business ethics. The Oxford Handbook of Business Ethics (2009). https://doi.org/10.1093/OXFORDHB/9780195307955.003.0002
204. Raji, I.D., Kumar, I.E., Horowitz, A., Selbst, A.: The fallacy of AI functionality. In: ACM International Conference Proceeding Series, pp. 959–972, Jun. 2022. https://doi.org/10.1145/3531146.3533158
205. Rahwan, I.: Society-in-the-loop: programming the algorithmic social contract. Ethics Inf. Technol. **20**(1), 5–14 (2018). https://doi.org/10.1007/s10676-017-9430-8
206. Dafoe, A.: AI governance: a research agenda, no. July 2017 (2017). https://doi.org/10.1176/ajp.134.8.aj1348938.
207. Truby, J., Brown, R.D., Ibrahim, I.A., Parellada, O.C.: A sandbox approach to regulating high-risk artificial intelligence applications. Eur. J. Risk Regul. **13**(2), 270–294 (2022). https://doi.org/10.1017/ERR.2021.52
208. Akpinar, N.-J., et al.: A sandbox tool to bias(Stress)-test fairness algorithms. ArXiv (2022). https://doi.org/10.48550/arxiv.2204.10233
209. Zinda, N.: Ethics auditing framework for trustworthy AI: lessons from the IT audit literature. In: Mökander J., Ziosi, M. (eds.) The 2021 Yearbook of the Digital Ethics Lab. Springer Cham (2021). https://doi.org/10.1007/978-3-031-09846-8
210. Mantelero, A.: AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. Comput. Law Secur. Rev. **34**(4), 754–772 (2018). https://doi.org/10.1016/j.clsr.2018.05.017
211. Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*, no. April, p. 22, 2018, [Online]. https://ainowinstitute.org/aiareport2018.pdf. Accessed 10 Feb 2023
212. Etzioni, A., Etzioni, O.: AI assisted ethics. Ethics Inf. Technol. **18**(2), 149–156 (2016). https://doi.org/10.1007/s10676-016-9400-6
213. Whittlestone, J., Clarke, S.: AI challenges for society and ethics. In: Bullock, J., Chen, Y.-C., Himmelreich, J., Hudson, V.M., Korinek, A., Young, M., Zhang, B. (eds.) The Oxford Handbook of AI Governance. Oxford University Press (2022). https://doi.org/10.1093/oxfordhb/9780197579329.013.3
214. Karan, M., Šnajder, j.: Preemptive toxic language detection in Wikipedia comments using thread-level context. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 129–134, Sep. 2019. https://doi.org/10.18653/V1/W19-3514
215. Gao, L., Huang, R.: Detecting online hate speech using context aware models. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 260–266, Nov. 2017, https://doi.org/10.26615/978-954-452-049-6_036
216. Delobelle, P., Tokpo, E.K., Calders, T., Berendt, B.: Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models. In: NAACL 2022—2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 1693–1706. (2022). https://doi.org/10.18653/V1/2022.NAACL-MAIN.122.
217. Nozza, D., Bianchi, F., Hovy, D.: HONEST: measuring hurtful sentence completion in language models. In: NAACL-HLT 2021—2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 2398–2406 (2021). https://doi.org/10.18653/V1/2021.NAACL-MAIN.191.
218. Costello, M., Hawdon, J., Bernatzky, C., Mendes, K.: Social group identity and perceptions of online hate. Sociol. Inq. **89**(3), 427–452 (2019). https://doi.org/10.1111/SOIN.12274
219. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N. A.: Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. In: NAACL 2022—2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 5884–5906 (2022). https://doi.org/10.18653/V1/2022.NAACL-MAIN.431
220. Kirk, H.R., Birhane, A., Vidgen, B., Derczynski, L.: Handling and Presenting Harmful Text in NLP Research. Findings of the Association for Computational Linguistics: EMNLP 2022 (2022). https://aclanthology.org/2022.findings-emnlp.35/
221. Welbl, J., et al.: Challenges in detoxifying language models. In: Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, pp. 2447–2469, Sep. 2021, https://doi.org/10.48550/arxiv.2109.07445

222. Rauh, M., et al.: Characteristics of harmful text: towards rigorous benchmarking of language models. ArXiv (2022). https://doi.org/10.48550/arxiv.2206.08325
223. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: CrowS-pairs: a challenge dataset for measuring social biases in masked language models. In: EMNLP 2020—2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1953–1967 (2020). https://doi.org/10.18653/V1/2020.EMNLP-MAIN.154
224. Rudinger, R.: GitHub - rudinger/winogender-schemas: Data for evaluating gender bias in coreference resolution systems. GitHub. https://github.com/rudinger/winogender-schemas (2019). Accessed 25 Jan 2023
225. Cihon, P., Kleinaltenkamp, M.J., Schuett, J., Baum, S.D.: AI Certification: advancing ethical practice by reducing information asymmetries. IEEE Trans. Technol. Soc. **2**(4), 200–209 (2021). https://doi.org/10.1109/tts.2021.3077595
226. Cihon, P., Schuett, J., Baum, S.D.: Corporate governance of artificial intelligence in the public interest. Information **12**(7), 1–30 (2021). https://doi.org/10.3390/info12070275
227. FDA.: Artificial intelligence and machine learning in software as a medical device. In: U.S. Food & Drug Administration, 2021. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (2021). Accessed 20 Jan 2023
228. Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, vol. 11, no. 21, pp. 375–385, Mar. 2021, https://doi.org/10.1145/3442188.3445901.
229. Mökander, J., Floridi, L.: Operationalising AI governance through ethics-based auditing: an industry case study. AI and Ethics (2022). https://doi.org/10.1007/s43681-022-00171-7
230. Mökander, J., Sheth, M., Gersbro-Sundler, M., Blomgren, P., Floridi, L.: Challenges and best practices in corporate AI governance: Lessons from the biopharmaceutical industry. Front. Comput. Sci. (2022). https://doi.org/10.3389/fcomp.2022.1068361
231. Smith, E.: Research design. In: H. Reis, H., Judd, C. (eds.) Handbook of Research Methods in Social and Personality Psychology, pp. 27–48 (2014) [Online]. https://doi.org/10.1017/CBO9780511996481.006
232. Sobieszek, A., Price, T.: Playing games with Ais: the limits of GPT-3 and similar large language models. Minds Mach. (Dordr) **32**(2), 341–364 (2022). https://doi.org/10.1007/s11023-022-09602-0
233. Floridi, L.: AI as Agency without Intelligence: on ChatGPT, large language models, and other generative models. SSRN. [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4358789 (2022). Accessed 13 Feb 2023
234. Reynolds, L., Ai, M., Ai, K., Mcdonell, K.: Prompt programming for large language models: beyond the few-shot paradigm. In: Conference on Human Factors in Computing Systems-Proceedings, May 2021, https://doi.org/10.1145/3411763.3451760
235. Hacking, I.: Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge University Press, Cambridge (1983)
236. Rorty, R.: Pragmatism as Anti-authoritarianism. Harvard University Press, Cambridge (2021)
237. Legg, C., Hookway, C.: Pragmatism. In Stanford encyclopedia of philosophy. PhilPapers (2020). https://plato.stanford.edu/entries/pragmatism/. Accessed 20 Feb 2020
238. Watson, D.S., Mökander, J.: In defense of sociotechnical pragmatism. In: Mazzi, F. (ed.) The 2022 Yearbook of the Digital Governance Research Group. Springer (2023). https://doi.org/10.1007/978-3-031-28678-0
239. Lee, M.S.A., Floridi, L., Singh, J.: "ormalising Trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. In: The 2021 Yearbook of the Digital Ethics Lab, pp. 157–182. Cham: Springer (2022). https://doi.org/10.1007/978-3-031-09846-8_11
240. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (Im)possibility of fairness. Commun ACM **64**(4), 136–143 (2021). https://doi.org/10.1145/3433949
241. Islam, G., Greenwood, M.: The metrics of ethics and the ethics of metrics. J. Bus. Ethics (2021). https://doi.org/10.1007/s10551-021-05004-x
242. Cugueró-Escofet, N., Rosanas, J.M.: The ethics of metrics: overcoming the dysfunctional effects of performance measurements through justice. J. Bus. Ethics **140**(4), 615–631 (2017). https://doi.org/10.1007/S10551-016-3049-2/TABLES/2
243. Boddington, P.: Towards a code of ethics for artificial intelligence. In: Artificial Intelligence: Foundations, Theory, and Algorithms. Switzerland: Springer Cham, (2017)
244. Minkkinen, M., Zimmer, M.P., Mäntymäki, M.: Towards Ecosystems for Responsible AI: Expectations, Agendas and Networks in EU Documents. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-85447-8
245. Schöppl, N., Taddeo, M., Floridi, L.: Ethics auditing: lessons from business ethics for ethics auditing of AI. In: Mökander J., Ziosi, M. (eds.) The 2021 Yearbook of the Digital Ethics Lab, pp. 209–227. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-09846-8
246. FDA.: Inspection classification database. U.S. Food & Drug Administration. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-classification-database (2022). Accessed 20 Jan 2023
247. British Safety Council.: About the British Safety Council. Website. https://www.britsafe.org/about-us/introducing-the-british-safety-council/about-the-british-safety-council/ (2023). Accessed 20 Jan 2023
248. Rainforest Alliance.: Our approach. Website. https://www.rainforest-alliance.org/approach/?_ga=2.137191288.953905227.1658139559-1130250530.1658139559 (2023). Accessed 20 Jan 2023
249. IAEA.: Quality management audits in nuclear medicine practices. IAEA Human Health Series, vol. 33. [Online]. http://www.iaea.org/Publications/index.html (2015). Accessed 20 Jan 2023
250. Duflo, E., Greenstone, M., Pande, R., Ryan, N.: Truth-telling by third-party auditors and the response of polluting firms: experimental evidence from India. Q. J. Econ. **128**(4), 1499–1545 (2013). https://doi.org/10.1093/QJE/QJT024
251. Tutt, A.: An FDA for Algorithms. Adm. Law. Rev. **69**(1), 83–123 (2017). https://doi.org/10.2139/ssrn.2747994
252. Carpenter, D.: Reputation and power: organizational image and pharmaceutical regulation at the FDA. In: Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA, pp. 1–802, (2014). https://doi.org/10.5860/choice.48-3548
253. Fraser, H.L., Bello y Villarino, J.-M.: Where residual risks reside: a comparative approach to Art 9(4) of the European Union's Proposed AI Regulation. SSRN Electron. J. (2021). https://doi.org/10.2139/SSRN.3960461
254. van Merwijk, C.: An AI defense-offense symmetry thesis. LessWrong. https://www.lesswrong.com/posts/dPe87urYGQPA4gDEp/an-ai-defense-offense-symmetry-thesis (2022). Accessed 20 Jan 2023
255. Du Sautoy, M.: The Creativity Code : Art and Innovation in the Age of A First US edition. Fourth Estate, Cambridge (2019)
256. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr) **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5

257. Frey, C.B.: The Technology Trap : Capital, Labor, and Power in the Age of Automation. Princeton University Press, Princeton (2019)
258. Mökander, J., Floridi, L.: From algorithmic accountability to digital governance. Nat. Mach. Intell. (2022). https://doi.org/10.1038/s42256-022-00504-5
259. Sloane, M.: The Algorithmic Auditing Trap. [Online]. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d (2021). Accessed 20 Jan 2023
260. Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., et al.: Adversarial training for high-stakes reliability. Adv. Neural Inf. Process. Syst. **35**, 9274–9286 (2022). arXiv:2205.01663
261. Keyes, O., Durbin, M., Hutson, J.: A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In: Conference on Human Factors in Computing Systems-Proceedings, May 2019, https://doi.org/10.1145/3290607.3310433
262. Mökander, J., Juneja, P., Watson, D.S., Floridi, L.: The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? Minds Mach (Dordr) (2022). https://doi.org/10.1007/s11023-022-09612-y
263. Epstein, Z., et al.: Turingbox: An experimental platform for the evaluation of AI systems. In: IJCAI International Joint Conference on Artificial Intelligence, vol. 2018-July, pp. 5826–5828 (2018). https://doi.org/10.24963/ijcai.2018/851
264. EPRS: A Governance Framework for Algorithmic Accountability and Transparency. European Parliamentary Research Service (2019). https://doi.org/10.2861/59990
265. EIU.: Staying ahead of the curve—the business case for responsible AI. The Economist Intelligence Unit. https://www.eiu.com/n/staying-ahead-of-the-curve-the-business-case-for-responsible-ai/, (2020). Accessed 7 Oct 2020
266. Mondal, S., Das, S., Vrana, V.G.: How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. Technologies **11**(2), 44 (2023). https://doi.org/10.3390/TECHNOLOGIES11020044
267. Muller, M., Chilton, L.B., Kantosalo, A., Lou Maher, M., Martin, C.P., Walsh, G.: GenAICHI: generative AI and HCI. In: Conference on Human Factors in Computing Systems-Proceedings, Apr. 2022, https://doi.org/10.1145/3491101.3503719
268. Rao, A.S.: Democratization of AI. A double-edged sword. Toward Data Science. https://towardsdatascience.com/democratization-of-ai-de155f0616b5 (2020). Accessed 22 Mar 2023
269. Salkind, N.J.: Encyclopedia of Research Design. SAGE, Los Angeles (2010)
270. Haas, P.J., Springer, J.F.: Applied policy research: concepts and cases. In: Garland Reference Library of Social Science ; v. 1051. New York: Garland Pub (1998)
271. Grant, M.J., Booth, A.: A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J **26**(2), 91–108 (2009). https://doi.org/10.1111/j.1471-1842.2009.00848.x
272. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: EASE '14 (2014). https://doi.org/10.1145/2601248.2601268.
273. Frey, B.B.: The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, vol. 4. SAGE Publications, Incorporated, Thousand Oaks (2018)
274. Adams, W.C.: Conducting semi-structured interviews. In: Handbook of Practical Program Evaluation: Fourth Edition, pp. 492–505. (2015). https://doi.org/10.1002/9781119171386.CH19.
275. Baldwin, R., Cave, M.: Understanding Regulation: Theory, Strategy, and Practice. Oxford University Press, Oxford (1999)
276. Vaswani, A., et al.: Attention is all you need. Adv Neural Inf Process Syst, vol. 2017-December, pp. 5999–6009, Jun. 2017. [Online]. https://arxiv.org/abs/1706.03762v5. Accessed 12 Apr 2023
277. Smith-Goodson, P.: "NVIDIA's New H100 GPU Smashes Artificial Intelligence Benchmarking Records. Forbes, 2022. [Online]. https://www.forbes.com/sites/moorinsights/2022/09/14/nvidias-new-h100-gpu-smashes-artificial-intelligence-benchmarking-records/?sh=5e8dca9ce728. Accessed 2 Apr 2023
278. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int J Comput Vis **115**(3), 211–252 (2015). https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16
279. Luccioni, A., Viviano, J.D.: What's in the Box? An analysis of undesirable content in the common crawl corpus. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 182–189 (2021). https://doi.org/10.18653/V1/2021.ACL-SHORT.24.
280. Han, X., et al.: Pre-trained models: past, present and future. AI Open **2**, 225–250 (2021). https://doi.org/10.1016/J.AIOPEN.2021.08.002
281. European Commission.: Ethics guidelines for trustworthy AI. AI HLEG, pp. 2–36, [Online]. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top (2019). Accessed 10 Feb 2023
282. Korbak, T., Elsahar, H., Kruszewski, G., Dymetman, M.: On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. ArXiv, Jun. 2022. [Online]. https://arxiv.org/abs/2206.00761v2. Accessed 2 Apr 2023
283. Min, S., et al.: Rethinking the role of demonstrations: what makes in-context learning work?, In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 11048–11064. [Online]. https://aclanthology.org/2022.emnlp-main.759 (2022). Accessed 2 Apr 2023
284. Alayrac, J.-B., et al.: Flamingo: a visual language model for few-shot learning. ArXiv (2022). https://doi.org/10.48550/arxiv.2204.14198
285. Zittrain, J.L.: The generative internet. Connect. Q. J. **13**(4), 75–118 (2014). https://doi.org/10.11610/CONNECTIONS.13.4.05
286. OpenAI.: Generative models. Blog post. https://openai.com/blog/generative-models/ (2016). Accessed 25 Jan 2023
287. NITS.: Red Team (Glossary). Computer Security Resource Center. https://csrc.nist.gov/glossary/term/red_team (2023). Accessed 2 Apr 2023
288. Bertuzzi, L.: AI Act: EU Parliament's crunch time on high-risk categorisation, prohibited practices. In: EURACTIV, 2023. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-parliaments-crunch-time-on-high-risk-categorisation-prohibited-practices/. Accessed 24 Mar 2023
289. AJL.: Algorithmic Justice League-Unmasking AI harms and biases. https://www.ajl.org/ (2023). Accessed 2 Apr 2023
290. Russell, S.J., Norvig, P.: Artificial Intelligence : A Modern Approach, 3rd edn. Pearson, New Delhi (2015)
291. Corning, P.A.: The re-emergence of emergence, and the causal role of synergy in emergent evolution. Synthese **185**(2), 295–317 (2010). https://doi.org/10.1007/s11229-010-9726-2
292. Molnar, C.: Interpretable machine learning. a guide for making black box models explainable.. Book, p. 247 (2021), [Online]. https://christophm.github.io/interpretable-ml-book. Accessed 10 Feb 2023
293. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. *Adv* Neural Inf. Process. Syst. **30** (2017)
294. Zang, S.: metaseq/projects/OPT/chronicles at main · facebookresearch/metaseq · GitHub. GitHub. https://github.com/facebookre

search/metaseq/tree/main/projects/OPT/chronicles (2022). Accessed 25 Jan 2023

295. Hubinger, E.: Relaxed adversarial training for inner alignment—AI Alignment Forum. In: Alignment Forum. https://www.alignmentforum.org/posts/9Dy5YRaoCxH9zuJqa/relaxed-adversarial-training-for-inner-alignment (2019). Accessed 20 Jan 2023
296. Weller, A.: Challenges for transparency. In: 2017 ICML Workshop on Human Interpretability in Machine (2017). https://openreview.net/forum?id=SJR9L5MQ-
297. Chasalow, K., Levy, K.: Representativeness in statistics, politics, and machine learning," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 77–89, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2101.03827.
298. Blodgett, S.L., Barocas III, S.H.D., Wallach, H.: Language (Technology) is power: a critical survey of 'Bias' in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476, Jul. 2020, https://doi.org/10.18653/V1/2020.ACL-MAIN.485.
299. Kirk, H.R., Vidgen, B., Röttger, P., Thrush, T., Hale, S. A.: Hatemoji: a test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In: NAACL 2022-2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 1352–1368 (2022), https://doi.org/10.18653/V1/2022.NAACL-MAIN.97.
300. Kumar, D., et al.: Designing toxic content classification for a diversity of perspectives. In: Proceedings of the Seventeenth Symposium on Usable Privacy and Security. https://data.esrg.stanford.edu/study/toxicity-perspectives (2021). Accessed 25 Jan 2023
301. Cantwell Smith, B.: The promise of artificial intelligence: reckoning and judgment. MIT Press, Cambridge (2019)
302. ForHumanity. Independent audit of AI systems (2023). https://forhumanity.center/independent-audit-of-ai-systems/. Accessed 12 Feb 2023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.