A Causal Framework for AI Regulation and Auditing

Lee Sharkey
Apollo Research
London, United Kingdom
lee@apolloresearch.ai

Clíodhna Ní Ghuidhir Apollo Research
Dan Braun Apollo Research
Jérémy Scheurer Apollo Research
Mikita Balesni Apollo Research

Lucius Bushnaq Apollo Research
Charlotte Stix Apollo Research
Marius Hobbhahn Apollo Research

Executive Summary

Artificial intelligence (AI) systems are poised to become deeply integrated into society. If developed responsibly, AI has potential to benefit humanity immensely. However, it also poses a range of risks, including risks of catastrophic accidents. It is crucial that we develop oversight mechanisms that prevent harm. This article outlines a framework for evaluating and auditing AI to provide assurance of responsible development and deployment, focusing on catastrophic risks. We argue that responsible AI development requires comprehensive auditing that is proportional to AI systems' capabilities and available affordances. This framework offers recommendations toward that goal and may be useful in the design of AI auditing and governance regimes.

Our main contributions are

- **A causal framework**: Our framework works backwards through the causal chain that leads to the *effects that AI systems have on the world* and discusses ways auditors may work toward assurances at each step in the chain.
- **Conceptual clarity**: We develop several distinctions that are useful in describing the chain of causality. Conceptual clarity should lead to better governance.
- **Highlighting the importance of AI systems' available affordances**: We identify a key node in the causal chain - the affordances available to AI systems - which may be useful in designing regulation. The affordances available to AI systems are the environmental resources and opportunities for affecting the world that are available to it, e.g. whether it has access to the internet. These determine which capabilities the system can currently exercise. They can be constrained through guardrails, staged deployment, prompt filtering, safety requirements for open sourcing, and effective security. One of our key policy recommendations is that proposals to change the affordances available to an AI system should undergo auditing.

We outline the causal chain leading to AI systems' effects on the world below, working backwards from the real world effects to inputs and determinants:

- **AI system behaviors** - The set of actions or outputs that a system actually produces and the context in which they occur (for example, the type of prompt that elicits the behavior).
- **Available affordances** - The environmental resources and opportunities for affecting the world that are available to an AI system.

- **Absolute capabilities and propensities** - The full set of potential behaviors that an AI system can exhibit and its tendency to exhibit them.
- **Mechanistic structure of the AI system during and after training** - The structure of the function that the AI system implements, comprising architecture, parameters, and inputs.
- **Learning** - The processes by which AI systems develop mechanistic structures that are able to exhibit intelligent-seeming behavior.
- **Effective compute and training data content** - The amount of compute used to train an AI system and the effectiveness of the algorithms used in training; and the content of the data used to train an AI system.
- **Security** - Adequate information security, physical security, and response protocols.
- **Deployment design** - The design decisions that determine how an AI system will be deployed, including who has access to what functions of the AI system and when they have access.
- **Training-experiment design** - The design decisions that determine the procedure by which an AI system is trained.
- **Governance and institutions** - The governance landscape in which AI training-experiment and security decisions are made, including of institutions, regulations, standards, and norms.

We identify and discuss five audit categories, each aiming to provide assurances on different determinants:

- AI system evaluations
- Security audits
- Deployment audits
- Training design audits
- Governance audits.

We highlight key research directions that will be useful for designing an effective AI auditing regime. High priority research questions include interpretability; predictive models of capabilities and alignment; structured access; and potential barriers to transparency of AI labs to regulators.

1 Introduction

AI has the potential to influence the lives of the public both positively [Jumper et al., 2021, Singhal et al., 2023] and negatively [Brundage et al., 2018, Anderljung and Hazell, 2023, Solaiman et al., 2023], including catastrophically [Hendrycks et al., 2023, Ngo et al., 2023]. As frontier AI systems become increasingly integrated into our lives and the economy, it grows ever more important to have assurances regarding the effects of these systems. Such assurances may be possible through auditing frontier AI systems and the processes involved in their development and deployment.

There are growing calls for regulations and audits of the social and technical processes that generate frontier AI systems. Auditing complements regulation; it helps to ensure that actors adhere to regulations and to identify risks before their realization. But the design and implementation of regulations for frontier AI are in their early days. How can we design a framework for auditing when most of the relevant regulations do not yet exist? Although we do not yet know which regulations will be implemented, many proposed measures already have broad support [Schuett et al., 2023] and therefore have a reasonable probability of being implemented. **Here, we discuss what a governance regime for frontier AI systems could look like and introduce a framework for the roles that auditors could play within it.** The framework focuses on extreme risks from frontier AI systems, rather than nearer term, smaller scale risks (such as bias or fairness) or risks from systems with narrower applications (such as lethal autonomous weapons). It emphasizes safety and assurance over other considerations, such as costs, political viability, or current technological barriers. Nevertheless, we believe that all regulations and auditor roles proposed here could and should be implemented in practice. It thus serves as a comprehensive menu of options for auditing in an AI governance regime and helps to identify areas for further research.

By 'auditing', we do not exclusively mean 'compliance auditing', where auditors ensure that auditees have complied with a well-defined set of acceptable practices, or where systems comply with a well-defined set of regulations. Although compliance audits are an ideal regime when possible, the science of AI safety is currently nascent enough that, in many areas of AI governance, compliance audits are at present inappropriate since standards are still emerging. Therefore, by 'auditing' we also include evaluations, which include risk assessments or assessments of whether AI systems possess certain dangerous capabilities. Evaluations may be one of the checks performed during auditing, though not necessarily compliance auditing.

Our work builds on similar previous work [Shevlane et al., 2023, Anderljung et al., 2023, Whittlestone and Clark, 2021, Mökander et al., 2023], though it differs in the following ways:

- **Causal framework**: Our framework begins with the target of our assurance efforts - *the effects that AI systems have on the world* - and works backwards through the causal process that leads to them. This provides an overarching framework for conceptualizing the governance of general-purpose AI systems. We hope that a framework will help to highlight potential regulatory blindspots and ensure comprehensiveness.
- **Conceptual engineering for clarity**: In developing the framework, we found we needed to use distinctions that we had not previously encountered. For instance, in section 2.2 we distinguish between absolute, contextual, and reachable capabilities, which we found useful for thinking about how AI systems interact with their environment and, hence, about regulations concerning their eventual effects on the world. We also reframe the focus from AI models to the slightly broader concept of AI systems (section 2.1).
- **Highlighting the importance of AI systems' available affordances**: One of the benefits of the technical focus of the framework is that it highlights a key concept, the affordances available to AI systems (defined in section 2.2) as a key variable for regulations, and serves as a unifying frame for many related concepts such as AI system deployment, open sourcing, access to code interpreters, guardrails, and more. The concept of available affordances adds important nuance to the 'training phase-deployment phase' distinction, a distinction that arguably formed the keystone of previous frameworks.

Our framework is agnostic to who should do the auditing. In general, however, we believe that *external* auditors, as well as auditors who are internal to AI labs, should be empowered to perform audits at all stages of AI system development. Decisions made prior to and during training ultimately affect the final AI system and therefore directly (and potentially radically) affect the public. Both

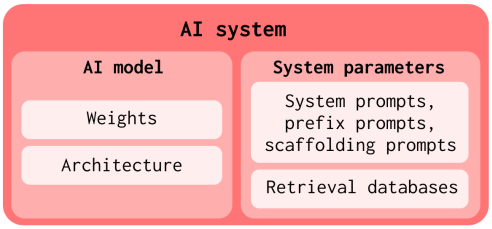

Figure 1: An AI system comprises an AI model, with its weights and architecture, as well as various other system parameters, including system prompts, prefix prompts, scaffolding prompts, and retrieval databases.

internal and external audits are necessary for ensuring public accountability and to ensure that the incentives of auditors are aligned with public interests [Raji et al., 2022, Mökander et al., 2023, Anderljung et al., forthcoming]. In addition to ensuring appropriate incentives, additional external auditing is also safer than relying solely on internal audits, since an absence of extensive auditing may reduce the possibility of discovering, and thus neutralizing, dangerous capabilities in AI systems.

2 Our framework: Auditing the determinants of AI systems' effects

2.1 Conceptualizing AI systems vs. AI models

The objects of our framework, *AI systems*, are a slight generalization of AI models. AI systems include not only the weights and architecture of the model, but also include a broader set of system parameters (Figure 1). These consist of retrieval databases and particular kinds of prompts. The reason for our expanded focus is that system parameters strongly influence the capability profile of AI systems. Therefore, systems with one set of system parameters may be safe to deploy whereas others are not. We discuss two main types of system parameter:

1. *System prompts, prefix prompts, and scaffolding*: System prompts and prefix prompts are prefixed to any user-generated prompts that are input to AI models (section 2.3.4). These prompts substantially modify system behavior. For instance: A model may not itself possess cyber offensive capabilities. But if part of its system prompt contains documentation for cyber offense software, then, through in-context learning (section 2.3.4), the system may exhibit dangerous cyber offense capabilities capabilities. This materially affects which systems are safe to deploy. If a lab wants to change the system prompt of a deployed model, it could change the capabilities profile enough to warrant repeated audits. Similarly, AI systems with few-shot prompts or chain-of-thought prompts (examples of *prefix prompts*) are more capable of answering questions than without those prompts that prefix their outputs. It is possible to arrange systems of models such that outputs of some models are programmatically input to other models (e.g. [Yao et al., 2023, Long, 2023, Gravitas, 2023]). These systems, which often involve using different *scaffolding prompts* in each of the models in the system, have different capability profiles than systems without scaffolding prompts.

2. *Retrieval databases*: Retrieval databases (section 2.3.4) may sometimes be considered parameters of the model [Borgeaud et al., 2022] and other times parameters of the AI system (of which the model is a part) [Lewis et al., 2020]. In both cases, the retrieval databases meaningfully influence the capabilities of the system, which may influence the outcomes of evaluations. We therefore expand the focus of what we consider important to evaluate in order to include all factors that influence the capabilities profile of the system.

AI models are still AI systems. They are simply a special case of AI systems that have only model parameters and no other AI system parameters. Given the relevance of learning and system parameters to capabilities and, hence, to auditing, we found it made most sense to place AI systems at the center of our framework, rather than AI models.

2.2 Conceptualizing AI system capabilities

In addition to shifting our focus from AI models to AI systems, we also develop some additional clarifying terminology regarding capabilities.

We use the following terms:

- **AI system behaviors.** The set of actions or outputs that a system actually produces and the context in which they occur (for example, the type of prompt that elicits the behavior). In this article, we exclusively mean behaviors that could be useful to the system or a user, rather than arbitrary or random behaviors that might be produced by an untrained system.
- **Available affordances.** The environmental resources and opportunities for influencing the world that are available to a system. We can exert control over the affordances available to a system through deployment decisions, training design decisions, guardrails, and security measures. As we'll see below, the available affordances determine which of a system's *absolute capabilities* are its *contextual capabilities and reachable capabilities* (Figure 3). Systems with greater available affordances have constant absolute capabilities but greater contextual and reachable capabilities (Figure 5).
- **Absolute capabilities.** The potential set of behaviors that a system could exhibit given any hypothetical set of available affordances (regardless of whether the system can reach them from its current context) (Figure 2).
  - ***Example***: If a trained system is saved in cold storage and not loaded, it is not contextually or reachably capable of behaving as a chat bot, even if it is absolutely capable of doing so. To become contextually or reachably capable of behaving as a chat bot, it must be loaded from storage and used for inference.
- **Contextual capabilities.** The potential set of behaviors that a system could exhibit right now given its current set of available affordances in its current environmental context (Figure 2).
  - ***Example***: A system may be absolutely capable of browsing the web, but if it does not have access to the internet (i.e., an available affordance), then it is not contextually capable of browsing the web.
- **Reachable capabilities.** The potential set of behaviors that a system could exhibit given its current set of available affordances (i.e. contextual capabilities) as well as all the affordances the system could eventually make available from its current environmental context. A system's reachable capabilities include and directly follow from a system's contextual capabilities (Figure 2).
  - ***Example***: A system, such as GPT4, may not be able to add two six-digit numbers. Therefore, six-digit addition is not within its contextual capabilities. However, if it can browse the web, it could navigate to a calculator app to successfully add the numbers. Therefore, six-digit addition is within its reachable capabilities.
- **System propensities.** The tendency of a system to express one behavior over another (Figure 3). Even though systems may be capable of a wide range of behaviors, they may have a tendency to express only particular behaviors. Just because a system may be capable of a dangerous behavior, it might not be inclined to exhibit it. It is therefore important that audits assess a system's propensities using 'alignment evaluations' [Shevlane et al., 2023].
  - ***Example***: Instead of responding to user requests to produce potentially harmful or discriminatory content, some language models, such as GPT-3.5, usually respond with a polite refusal to produce such content. This happens even though the system is capable of producing such content, as demonstrated when the system is 'jailbroken'. We say that such systems have a propensity not to produce harmful or offensive content.

It is worth noting that the 'capabilities' of a system are multidimensional in a similar sense that 'intelligence' is multidimensional. A 'less capable' system might have learned some abilities that a 'more capable' system has not. However, when comparing two general cognitive systems, we rely on the intuition that a 'more capable' system has learned almost all and more of the abilities that 'less capable' has learned.

It is also worth noting that training has an effect on all classes of capabilities (absolute, contextual, reachable), but some more than others. For example, if a large language model is being trained offline

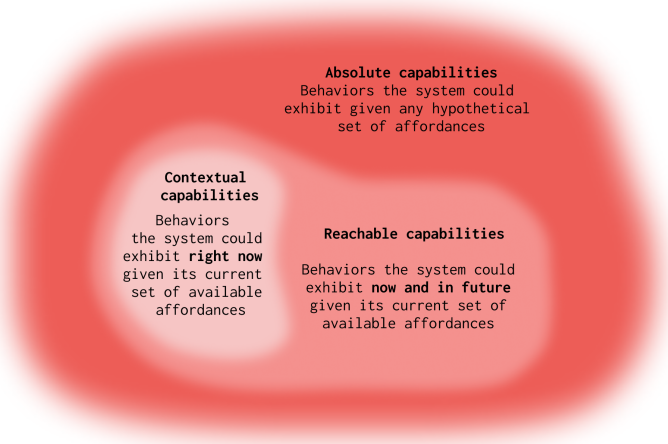

Figure 2: The relationship between the sets of potential behaviors defined by absolute capabilities, reachable capabilities, and contextual capabilities.

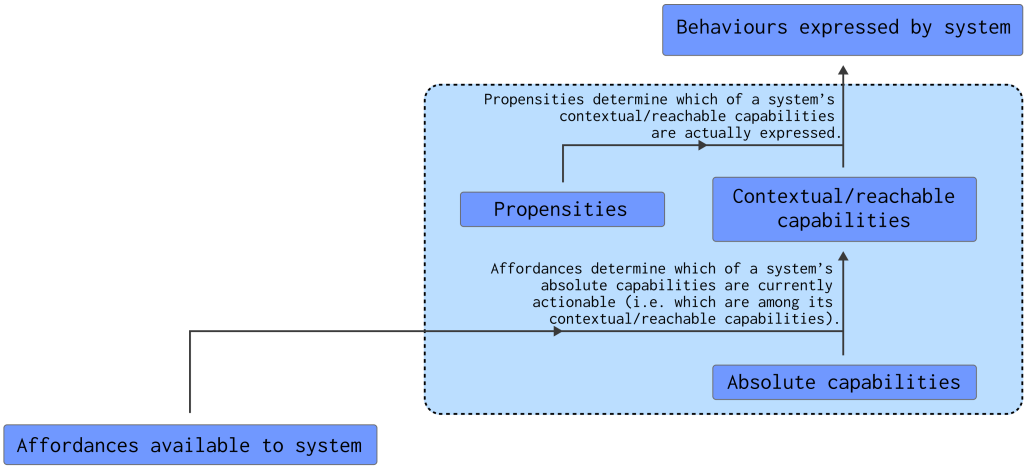

Figure 3: The relationship between an AI system's capabilities, propensities, affordances, and behaviors.

on a large fixed corpus of text (i.e. it is not learning online by, for instance, browsing the web and updating its weights frequently during browsing), then during training its contextual capabilities remain quite limited - they are limited only to outputting a prediction for the current minibatch of data. The system's reachable capabilities remain similarly limited. However, its absolute capabilities may increase a lot throughout the training process. This is because absolute capabilities do not depend on a system's current set of available affordances. It is important to design training-experiments that do not introduce risks from excessive absolute capabilities (such as a training-experiment that scales an AI system 1000x larger before understanding similar systems' capabilities at smaller scales) or excessive affordances (such as giving an AI system root access to a supercomputing cluster or unconstrained access to the internet).

2.3 The determinants of AI systems' effects on the world

The ultimate purpose of auditing for AI safety is to reduce the risk of AI systems having damaging effects on the world. In particular, the role of auditing is to identify sources of risk before the damage materializes. It is therefore important to audit the *upstream determinants of AI systems' effects on the world* (Figure 4).

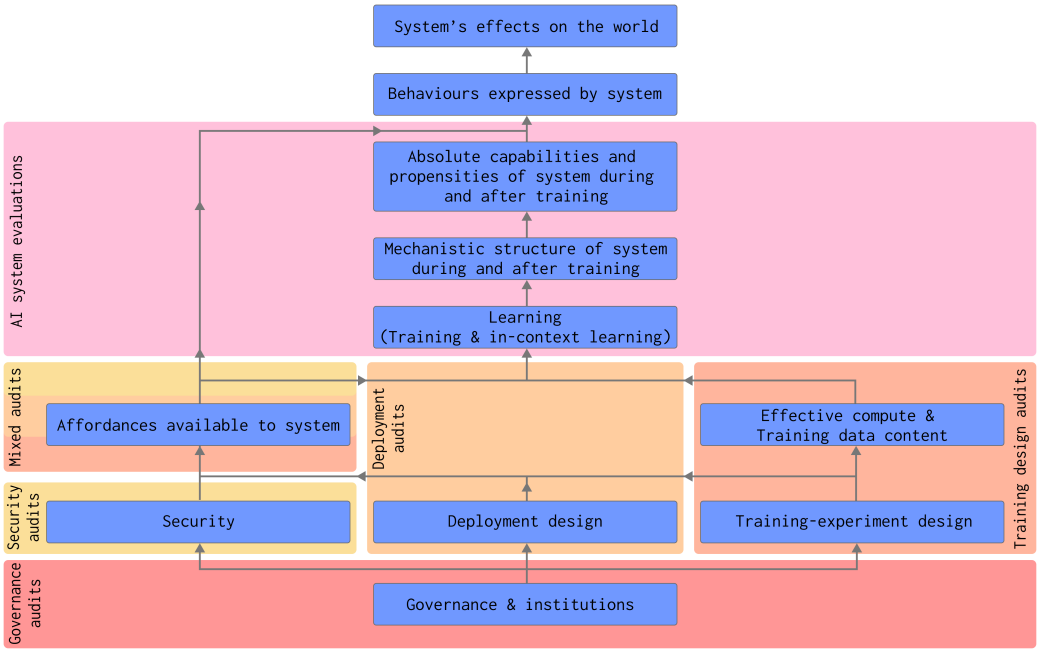

Figure 4: Determinants of AI system's effects on the world and the types of auditing that act on them.

AI systems' effects on the world are determined by the behaviors they express, which are in turn determined by:

1. Affordances available to the system
2. Absolute capabilities and propensities of the system during and after training
3. Mechanistic structure of the system during and after training
4. Learning
5. Effective compute and training data content
6. Security
7. Deployment design
8. Training-experiment design
9. Governance and institutions

We identify five main kinds of audit, each aiming to provide assurances on different determinants (Figure 4). These are:

1. AI system evaluations
2. Security audits
3. Deployment audits
4. Training design audits
5. Governance audits.

In the sections below, we consider each determinant in turn and explore auditors' potential roles with respect to each.

2.3.1 Affordances available to a system

The affordances available to an AI system determine which of its absolute capabilities are among its contextual and reachable capabilities. Systems with greater available affordances have constant

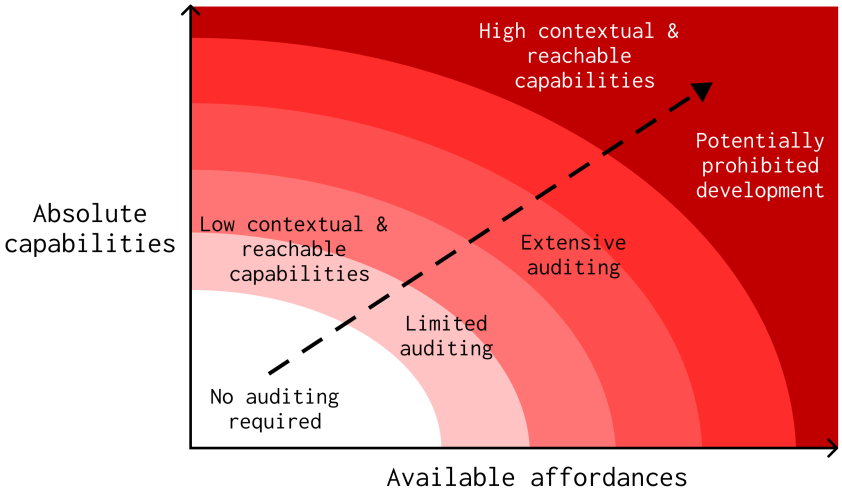

Figure 5: The relationship between absolute capabilities, affordances, contextual and reachable capabilities, and the level of auditing warranted. Absolute capabilities and available affordances are orthogonal. As either increase, the level of auditing required also increases.

absolute capabilities but greater contextual and reachable capabilities (Figure 5). The potential harm posed by an AI system can thus increase with increasing available affordances. **One of our key recommendations is that decisions that expand the affordances available to AI systems should be subject to auditing.**

The affordances available to AI systems are determined by a number of factors, which we'll look at in turn. They include:

1. Extent of system distribution
2. Actuators, interfaces, and scaffolding
3. Online vs offline training
4. Guardrails

Extent of system distribution This is the number of people the AI system interacts with and the extent to which it is integrated into systems in the wider world. This includes factors such as deployment, open sourcing, and other forms of AI system proliferation.

'Extent of system distribution' closely reflects the training vs. deployment distinction emphasized in other auditing frameworks [Shevlane et al., 2023, Mökander et al., 2023]. However, it is an imperfect reflection, since AI systems may potentially have access to dangerous levels of affordances during training, for instance during online training. The 'training vs deployment' distinction is only relevant for risk because it reflects, albeit imperfectly, the affordances available to a system. Thus, while not entirely disposing of the training versus deployment distinction, available affordances are more important in our framework because they are more direct causes of risk.

When a developer proposes deployment of an AI system, it should be subject to a deployment audit because it involves a change to the system's available affordances. Sudden jumps in the extent of system distribution available to systems represent sudden jumps in contextual and reachable capabilities. This may lead to unexpected adverse outcomes. For example, if a system with the capability to instruct users on how to build bioweapons is open sourced, then it may be usable by malicious actors to devastating effect. Staged deployment, first to trusted AI system evaluations researchers (for instance) and finally to the broader public, may help identify and mitigate these risks while access is expanded. In other words, evaluations researchers may help estimate the absolute capabilities of the system before very large affordances (and hence potentially large contextual and reachable capabilities) are made available to them. At each stage, *deployment audits* (Figure 4) should be required before a new deployment stage begins. In a similar way, AI systems should not be

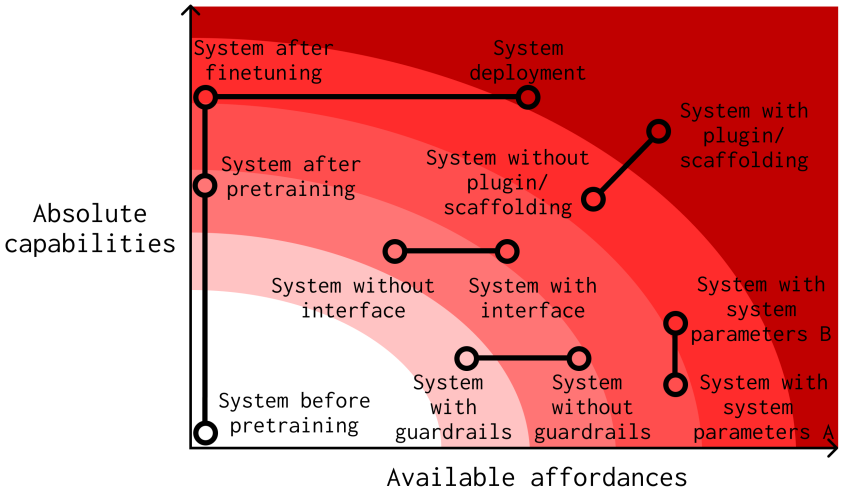

Figure 6: The effects of different types of training and available affordances on capabilities. While changes in available affordances increase contextual and reachable capabilities, they don't change absolute capabilities.

open-sourced unless it can be demonstrated (through a deployment audit) that the risks are sufficiently small.

Besides intentional deployment, another way that the extent of system distribution may increase is unintentional proliferation, e.g. through leaked or stolen model weights [Heim, 2022] or, in highly capable future AI systems, the system exfiltrating itself. *Security audits* (Figure 4) should help reduce the risks from system theft or autoexfiltration. The weights of highly absolutely capable systems should also be tracked, such that the locations of all copies are known and secured.

Expensive AI systems with extensive absolute capabilities are likely to be targets of theft. They are also more likely than less capable systems to be capable of autoexfiltration. Therefore, the risk of proliferation may increase as absolute capabilities increase, and therefore also increases the risk of unintended levels of affordances becoming available to the system. This risk means that absolute capabilities and available affordances may become positively correlated at high levels of absolute capabilities. Once systems proliferate widely, protecting society from potential harms becomes exceptionally, if not impossibly, difficult [Anderljung and Hazell, 2023]. We should therefore be careful about developing highly absolutely capable systems at all.

Actuators, interfaces, and scaffolding AI system outputs can be input to specialized interfaces. These may be physical, such as monitors (and the people looking at them) and robotic actuators (e.g. [Boiko et al., 2023, Brohan et al., 2022]). But they may also be interfaces to software systems, such as code interpreters [OpenAI, 2023a] or actions in games [Reed et al., 2022, Wang et al., 2023], or even as inputs to other language models.

Systems can be furnished with plugins or tools [Bubeck et al., 2023], where systems are given access to interfaces and are provided with prompts (or potentially fine-tuning [Schick et al., 2023]) that teach them how to use those interfaces. Plugins serve to facilitate a system's interactions with a wider range of interfaces and actuators. For example, an AI system without the right plugin may not easily be able to interact with the internet. But when given access to a browsing interface and prompts that inform the system on how to use the interface, the system may be able to browse the internet effectively.

Similar to plugins is scaffolding. *Scaffolding programs* involve passing the outputs of one AI system to the inputs of other system instances, sometimes with additional scaffolding prompts that guide models on how to use the outputs of the other models. Scaffolding programs influence the available affordances of AI systems used within the scaffolding, whereas *scaffolding prompts* shape their capabilities. One example of scaffolding is 'Tree of Thought' scaffolding [Yao et al., 2023], which uses multiple language models scaffolded together to answer questions; one AI system produces multiple candidate intermediate steps to answer a problem, and a second system evaluates those

candidates. The scaffolded systems move closer to answer the question by advancing along only the most promising intermediate steps. Other frameworks exist, such as LangChain or AutoGPT, which give systems the ability to behave like autonomous agents. Today's AI systems do not appear to be able to make maximum use of scaffolding due to reliability issues. Nevertheless, near-term future systems with marginally greater absolute capabilities might be reliable enough to make proficient use of these affordances, which has potential to vastly increase their contextual and reachable capabilities.

AI systems may be given different actuators, interfaces, or scaffolding at different points in training and deployment. This means that both deployment audits and *training design audits* (Figure 4) should be used in their oversight. Security audits may also be involved in ensuring that AI systems do not have access to more actuators or interfaces than intended.

Online vs offline training In *offline training*, the training dataset is collected and fixed prior to training. Systems have limited affordances available to them during offline training since their outputs mostly do not affect anything in the world. Nor can they accrue more affordances during training. Therefore the system's contextual and reachable capabilities remain roughly constant, although their absolute capabilities increase.

By contrast, *online training* involves training on data that are collected during training. The training data result from the outputs of the AI system that is being trained and their interaction with its (optional) environment. Systems trained online usually interact with an environment, and those interactions require that the environment provide the system with some affordances. If that environment is the real world, such as a chatbot interacting with humans, it elevates risks. Online training therefore shares similarities with deployment: In both cases, AI systems have extensive affordances available and may have real world consequences. This makes online training potentially riskier. Decisions regarding online training should therefore be audited during training design audits.

Extent of guardrails We can potentially limit the affordances available to AI systems by introducing guardrails. For example, attempts were made - with qualified success [Rando et al., 2022] - to prevent the AI system Stable Diffusion from outputting sexually explicit images by applying a safety filter to the output of the image generation system. In the language model case, it may be possible to curtail certain kinds of capabilities by e.g. filtering their inputs (using another language model) such that any prompt that has the system do chain-of-thought reasoning is prevented from being input to the AI system. It may also be possible to curtail the acquisition of new affordances through guardrails. An advanced AutoGPT-like system could, for instance, have guardrails that require the AI system to request human permission to access the internet; to use financial resources; to create new system instances; or to run code. The job of ensuring that appropriate guardrails are in place should fall on either deployment audits or, if guardrails will be imposed during training, training design audits.

Summary of recommendations regarding affordances available to AI systems

- Proposed changes in the affordances available to a system (including changes to the extent of a system's distribution, online/offline training, actuators, interfaces, plugins, scaffolding, and guardrails) should undergo auditing, including conduct risk assessments, scenario planning, and evaluations.
- AI systems' deployment should be staged such that distribution increases in the next stage only if it is deemed safe.
- Model parameters should not be open sourced unless they can be demonstrated to be safe.
- All copies of highly absolutely capable models should be tracked and secured.
- Guardrails should be in place to constrain the affordances available to AI systems.

2.3.2 Absolute capabilities and propensities of a system

One of the goals of *AI system evaluations* (Figure 4) is to assess whether a system will exhibit any dangerous behaviors when certain affordances are made available to it. When systems are deployed, users could for instance provide the system with novel scaffolding. For sufficiently absolutely capable systems, these affordances may give the system potentially dangerous reachable capabilities. By looking for these dangerous capabilities in more controlled environments, AI systems evaluations can help us better predict the risks posed by deployment.

When evaluating a system, we should assess whether the system possesses certain capabilities ('dangerous capability evaluations') and whether it has the propensity to exhibit them ('alignment evaluations') [Shevlane et al., 2023]. These should be done both during and after training to ensure system developers and auditors understand the current capabilities and propensities at any training checkpoint and to respond to trends in how systems are being used. Certain kinds of training algorithms may result in systems with different absolute capability and propensity profiles; for instance, systems trained with RL algorithms may potentially result in more dangerous systems than systems trained with other algorithms [Turner et al., 2019].

Attempts to elicit dangerous behaviors in a controlled setting may involve prompt engineering [Reynolds and McDonell, 2021] or prompt optimization [Shin et al., 2020, Wen et al., 2023, Jones et al., 2023]. Fine-tuning methods, such as prefix tuning [Li and Liang, 2021] or standard fine-tuning (for instance, using RLHF [Christiano et al., 2017]), may serve as approximate upper bounds on what can be achieved by prompting alone. Evaluating the capabilities of fine-tuned systems can also be used to inform the decision of whether to deploy fine-tuning access to users.

It is likely the case that dangerous capability evaluations are easier than alignment evaluations; to demonstrate that an AI system can exhibit the dangerous capability in any setting; to demonstrate an AI system's propensity to exhibit a dangerous capability, it must be shown whether or not the system exhibits it in all (or at least a representative sample) of settings. This seems much harder to do, since the space of possible settings is very large.

It is important to be clear that, like AI system training, dangerous capability evaluations and alignment evaluations can be 'gain-of-function' work, where dangerous capabilities are elicited in controlled settings. As systems become closer to being able to, for instance, autonomously self-replicate or deceive evaluators, the risk of these experiments increases. At that point, auditors must therefore be careful to avoid introducing new risks through their work. Eventually, before such systems are developed, there should be

1. Risk assessments prior to certain gain-of-function experiments to ensure that the risks are worth the benefits.
2. Requirement of official certification to perform certain gain-of-function work, reducing risks from irresponsible or underqualified actors.
3. Information controls. Given the risk that information obtained by gain-of-function research may proliferate, there should be controls on its reporting. Auditors should be able to report gain-of-function results to regulators, who may have regulatory authority on decisions whether to prevent or reduce system deployment, and not necessarily to AI system developers or the broader scientific community.

To avoid hampering the field in its early stages, it seems sensible to implement these requirements only for models considerably larger (and therefore riskier) than today's. This said, as AI systems are trained, their absolute capabilities increase and highly absolutely capable systems may be able to make use of even very limited available affordances. For example, suppose a dangerous system had no access to the outside world other than through a single terminal to interact with researchers; if the system has sufficient absolute capabilities, it may be able to use its limited set of available affordances to manipulate the auditors and developers into giving it access to even greater affordances, such as internet access, which may be dangerous. In other words, if a system has sufficient absolute capabilities, its limited set of available affordances may nevertheless give it dangerously large reachable capabilities even though its contextual capabilities are small. Therefore, as absolute capabilities increase, gain of function research becomes riskier, so there should be increasing restrictions on both the development of such absolutely capable systems as well as on gain-of-function research that involves such systems.

It is important that enforceable action plans are triggered following AI system evaluations that reveal danger. Unless the evaluations lead to concrete actions that mitigate risk, they may not be worth the additional risks they create. Without enforceable protocols that describe what will be done following specific concerning evaluations, such as the prevention of deployment or the pausing of training, then there is little purpose in doing these evaluations at all. Some action plans could be implemented in code, such that training does not proceed if certain evaluations reveal cause for concern.

Summary of proposed changes regarding absolute capabilities and propensities of AI systems

- Evaluate dangerous capabilities and propensities continually throughout and after training.
- When such experiments involve extremely capable AI systems, auditors should require certification to perform gain-of-function experiments, similar to AI developers; risk assessments prior to gain-of-function experiments should be required; and information controls on reporting gain-of-function results should be implemented.
- Have enforceable action plans following concerning evaluations.

2.3.3 Mechanistic structure of a system

The AI systems under discussion are neural networks that have learned complicated internal algorithms that implement highly capable input-output functions. The *mechanistic structure* of an AI system consists of the algorithms it implements internally. The field of *mechanistic interpretability* aims to characterize AI systems' mechanistic structure. Mechanistic structure is determined not only by internal algorithms implemented by the model (i.e. the architecture and network parameters), but also by its system parameters (e.g. system prompts, prefix prompt, or retrieval database).

Evaluations of AI system behavior alone will, unfortunately, shed limited light on a system's absolute capabilities; they can only show that certain behaviors are among a system's absolute capabilities but cannot conclusively show that certain behaviors are not. If we understood a system's mechanistic structure we could predict what it will do in particular hypothetical situations. For instance, if we observed that a system had a mechanistic structure such that it possessed no knowledge about dangerous pathogens, we would be able to predict that it wouldn't be capable of giving instructions on how to manufacture a bioweapon when asked. Understanding a system's mechanistic structure should therefore give us a better sense of its absolute capabilities, which will give us a much better understanding of a system's reachable and contextual capabilities in arbitrary settings and therefore let us make predictions about system behavior that are relevant to safety. It should also give us a better understanding of systems' propensities, we would be able to make wider inferences about what systems would tend to do in different contexts.

At present, mechanistic interpretability, the field of research that would let us understand systems' mechanistic structure, is in its infancy. Current interpretability methods seem inadequate for making confident claims about how AI systems might behave outside of the evaluation setting [Levinstein and Herrmann, 2023]. It will therefore be important to develop better interpretability methods and incorporate them into AI system evaluations. Nevertheless, in advance of adequate interpretability methods being found, it will be important to develop forms of structured access [Trask et al., 2023, Shevlane, 2022] for external auditors that permit the kinds of interpretability research that will be necessary to gain assurances of AI systems' safety.

Summary of recommendations regarding the mechanistic structure of AI systems

- Do not overstate the guarantees about an AI system's absolute capabilities and propensities that can be achieved through behavioral evaluations alone.
- Incorporate interpretability into capability and alignment evaluations as soon as possible.
- Develop forms of structured access for external auditors to enable necessary research and evaluations.

2.3.4 Learning

Learning is the process by which AI systems develop mechanistic structures that are able to exhibit intelligent-seeming (sometimes dangerous) behavior. Here, this includes both training of model parameters by gradient descent (i.e. pretraining and fine-tuning) and other forms of optimization such as in-context learning [von Oswald et al., 2023, Xie et al., 2022].

Pretraining and fine-tuning The typical development pipeline for general-purpose AI systems involves a long offline 'pretraining' phase on a large corpus of data. Pretraining need not be exclusively offline, but today usually is. Pretraining is often followed by multiple 'fine-tuning' phases. There are diverse approaches to fine-tuning, which have a wide variety of effects on systems' absolute capabilities including (but not limited to):

- Instruction fine-tuning [Wei et al., 2022a], RLHF [Christiano et al., 2017], Constitutional AI [Bai et al., 2022] drastically change a system's absolute capabilities and propensities;
- Iterative hallucination reduction, as in GPT-4 [OpenAI, 2023b];
- Different training objectives that increase absolute capabilities on downstream metrics, such as UL2 training [Tay et al., 2022];
- Adding new parameters to a system and fine-tuning on particular tasks, such as in LoRA fine-tuning or prefix-tuning [Hu et al., 2021, Li and Liang, 2021];
- Fine-tuning two independently trained models so they operate as one model, such as Flamingo, which combines a vision model with a language model [Alayrac et al., 2022];
- Enabling a model to use particular kinds of retrieval databases [Borgeaud et al., 2022, Wu et al., 2022];

The main differences between pretraining and fine-tuning is that pretraining comes first and the number of updates applied to the model parameters is usually much larger than in fine-tuning. The relevant aspect of pretraining vs fine-tuning with respect to auditing is how much they contribute to system absolute capabilities and propensities. Audits should be required according to the level of absolute capability a particular training phase may add, regardless of number of updates. Although some fine-tuning updates may have only small changes to the absolute capabilities of a system, some fine-tuning approaches may contribute significant absolute capabilities with only small numbers of updates. For example, using the UL2R mixture of training objectives to finetune a language model, it is possible to drastically improve performance on downstream tasks using only 0.1% to 1% of the pretraining computation costs [Tay et al., 2022]. Fine-tuned systems may therefore necessitate significant additional auditing.

AI systems evaluations during training will be important for developing good predictive models of capabilities, which we currently lack [Srivastava et al., 2023, Wei et al., 2022b, Ganguli et al., 2022]. By modeling patterns and phase changes in a system's capability profile, we would like to be able to predict when the capacity for certain dangerous behaviors will emerge. This may let us avoid training AI systems capable of particularly dangerous behaviors, such as strategic deception, which may render other evaluations useless or have catastrophic outcomes.

Learning from prompts and retrieval databases A remarkable property of large language models is their capacity for in-context learning [Brown et al., 2020, Olsson et al., 2022, Raventós et al., 2023, von Oswald et al., 2023, Wei et al., 2023, Xie et al., 2022]. For example, when provided with in-context demonstrations of how to perform a task, a system that can't solve a task zero-shot may become able to solve the task after being provided with a few demonstrations [Brown et al., 2020]. The absolute capabilities of an AI system with one prompt are different than with a different prompt; this is because the structure of the function they implement is different thanks to different system parameters, even though the weights of the model stay constant.

While the ability to learn new capabilities in-context is one of the most useful properties of language models, it is also one of the main sources of risk. For instance, a system may in-context-learn new cyber offense skills from documentation for cyber offense software. As discussed in section 2.3.1, plugins and scaffolding often employ prompts to teach AI systems how to use software interfaces or operate within scaffolding programs, giving them very different capability profiles. Therefore, AI systems evaluations are needed to test the extent to which systems can learn dangerous capabilities in-context. This includes the ability to behave as an autonomous agent using scaffolding such as LangChain [Langchain, 2023] or AutoGPT [Gravitas, 2023].

As introduced in section 2.1 the other kind of system parameter is retrieval databases, which make use of or influence in-context learning [Borgeaud et al., 2022, Izacard et al., 2022, Zhong et al., 2022, Karpukhin et al., 2020, Lewis et al., 2020, Guu et al., 2020]. As with prompts, changes to a model's retrieval database could change the capabilities profile of the model. To reuse the same example, adding information about cyber offense to a model's retrieval database may furnish a model with novel cyber offensive capabilities. Changes to retrieval databases should necessitate significant auditing.

There are multiple kinds of retrieval methods; not all have the same auditing requirements. Most kinds use an embedding model to vectorize a corpus of data, such as Wikipedia or the entire training dataset,

and then look up useful vectors in this database using vector similarity metrics. From that point, different methods use different approaches. For instance, Lewis et al. [2020] insert the looked-up data into the prompt so that the language model can use the extra information in the prompt to better predict the next token. Another method, RETRO [Borgeaud et al., 2022] instead incorporates the database vectors using cross-attention, which the network learns through finetuning. The retrieval database in Lewis et al. [2020] may be best thought of as system parameters, whereas the database in Borgeaud et al. [2022] may be best thought of as model parameters. In both cases, changes to retrieval databases should warrant AI system evaluations, since they may change systems' absolute capabilities. A special case of retrieval is 'memorizing transformers' [Wu et al., 2022], where the embedding model is the language model itself at previous timesteps (i.e. the retrieval database consists of the hidden activations of the AI system for earlier parts of the dataset during training). This is functionally similar to giving the system a much larger context window. Retrieval databases of memorizing transformers are thus constructed 'online', whereas retrieval databases as in RETRO [Borgeaud et al., 2022] are constructed 'offline'. Similar to training data, it will be more challenging to audit retrieval databases that are constructed online since auditing must be performed constantly; by contrast, we can audit an offline-constructed retrieval database once.

Summary of recommendations regarding learning

- For each new experiment, require audits in proportion to expected capability increases from pretraining or fine-tuning.
- Filter prompts and retrieval databases to avoid AI systems learning potentially dangerous capabilities using in-context learning.
- When system parameters, such as retrieval databases, are changed, the AI system should undergo renewed auditing.

2.3.5 Effective compute and training data content

Changes in a system's mechanistic structure that increase absolute capabilities necessitate further auditing. We identify the main inputs to a system's absolute capabilities as 1) Effective compute and 2) Training data content.

Effective compute Vast leaps in the absolute capabilities of AI systems in recent years have been, to a large extent, driven by similarly vast increases in the size of systems, the amount of data, and the amount of computation used to train them [Kaplan et al., 2020, Hoffmann et al., 2022, Sevilla et al., 2022]. There has also been significant *algorithmic progress*, which reduces the amount of compute needed to obtain the same performance [Erdil and Besiroglu, 2023, Tucker et al., 2020]. This means that even if we held compute constant, algorithmic progress would mean that *effective compute* would continue to increase, where effective compute is the product of the amount of compute used and the efficiency of how that compute is used. Systems trained with additional effective compute leads to increased absolute capabilities, introducing further risks, which should necessitate additional auditing.

Like absolute capabilities, effective compute is difficult to measure. This makes effective compute impractical for use as a policy lever. Instead, regulators, and hence auditors, will likely focus on correlates: Amount of compute and algorithmic progress.

- **Amount of compute** Given that AI systems that undergo additional training may have increased absolute capabilities, they should require additional auditing. Unfortunately, even though the absolute amount of computation used in training-experiments correlates with a system's absolute capabilities, it is not possible to use it to predict accurately when particular capabilities will emerge [Srivastava et al., 2023, Wei et al., 2022b, Ganguli et al., 2022]. We want to avoid risky scenarios where systems trained with increasing amounts of compute suddenly become able to exhibit certain dangerous behavior. We may be able to do this by evaluating systems trained with a lower level of compute to understand their capabilities profiles before moving to systems trained with slightly more. Risk assessments should ensure that slightly smaller systems have undergone adequate auditing before permission is given to develop larger systems. If compute thresholds exist, auditors may ensure that training-experiment designs and the experiments themselves do not exceed permitted thresholds.

- **Algorithmic progress** AI is a dual-use technology; it has great potential for both good and harm. Due to its dual-use nature, widespread access to large amounts of effective compute could result in proliferation of dangerous misuse risks (such as systems capable of automating the discovery of zero-day exploits on behalf of bad actors) or accident risks (such as systems capable of autonomous exfiltration and replication). Algorithmic progress is one of the key inputs to effective compute. Thus, when tasked with governing effective compute, policymakers are faced with a challenge: Algorithmic improvements are often published openly. But future publications may have unpredictable effects on the amount of effective compute available to all actors. Therefore, the standard publication norm or openness may therefore unintentionally provide dangerous actors with increased effective compute. Policymakers may therefore consider it necessary from a security perspective to implement publication controls, such as requiring pre-publication risk assessments, in order to prevent undesirable actors gaining access to potentially dangerous amounts of effective compute. Such publication controls for dual-use technologies have precedent; they are the norm in nuclear technology, for instance [Wasil et al., 2023]. It may not be possible to rely on lab self-regulation, since labs are incentivised to publish openly in order to garner prestige or to attract the best researchers. Regulation would therefore likely be required to ensure that pre-publication risk assessments take national security into account sufficiently. Assessing risks of publication may require significant independent technical expertise, a role which regulators may either have in-house or could be drawn from auditing organizations.

  In addition to facilitating pre-publication risk assessments, auditing organizations may also serve as monitors of algorithmic progress within labs, since doing so would require access to frontier AI systems in order to evaluate them, adequate technical capacity, and independence from other incentives. Between AI labs, regulators, and auditors, it may therefore be the case that auditors are the best positioned actors to perform this function. However, currently no consensus metric of algorithmic progress exists and further research is required to identify metrics that are predictive of capability levels. Building a 'science of evals' and designing metrics that are more predictive of capabilities than compute should be priority research areas.

Training data content AI system capabilities emerge as a result of learning from the training data. This means that we can exert some control over system behavior by controlling the data it is trained on. For instance, we can control AI systems' absolute capabilities [Gunasekar et al., 2023, Eldan and Li, 2023, Raffel et al., 2023] or ease of alignment [Korbak et al., 2023] by controlling for training data quality or training only on certain subsets of data [Chan et al., 2022]. Similarly, we may be able to avoid certain dangerous capabilities from emerging at all by carefully curating an AI system's training data. For instance, if we want to ensure that a system cannot easily tell users how to synthesize illicit substances (even if 'jailbroken' [Zou et al., 2023]), then we should remove data related to the synthesis of those substances from the training data; similarly, if we wish to limit a system's ability to exfiltrate itself, we should consider removing cybersecurity-related data from its training corpus. However, filtering data is unlikely to yield strong guarantees about AI system capabilities because AI systems can generalize. For instance, a system might be able to give instructions on how to synthesize dangerous chemicals using its knowledge of how to synthesize similar, safe compounds. In some instances, data may therefore be inherently dual-use. Nevertheless, figuring something out for oneself is harder than being taught; we can therefore potentially make some behaviors harder to obtain than they otherwise would be by filtering data. External and internal auditors should evaluate training data pertaining to potentially dangerous behaviors and filter as appropriate.

When auditing of data should occur will depend on certain aspects of training. In offline training, the training data are pre-collected prior to training. By contrast, online training involves data that are collected during training, often as a result of interactions with an environment. Auditing the training data of an AI system trained offline can be done by filtering its corpus for sensitive data. This step may be expensive, but only need be done once, prior to beginning training. Filtering data during online training may be more expensive since it must be performed continually, but we otherwise have few safety assurances over online training data content. Iterative retraining has properties of both offline and online training; data content should be audited prior to each bout of training.

Summary of recommendations regarding effective compute and training data content

- For each new experiment, require audits in proportion to expected capability increases from additional effective compute or different training data content.
- Conduct risk assessments of slightly smaller AI systems before approval is given to develop a larger system.
- Place strict controls on training-experiments that use above a certain level of effective compute.
- Implement national security-focused publication controls on research related to AI capabilities.
- Auditors should be able to evaluate training data.
- Filter training data for potentially dangerous or sensitive content.

2.3.6 Security

Preventing AI system theft and espionage Proper security measures are crucial for controlling the affordances available to systems. Adequate security helps avoid several risky scenarios such as proliferation of the system by malicious actors (including advanced threat actors such as hostile governments) or the system self-proliferating. Any organization interacting with AI systems and the computer infrastructure that they run on should implement strict cyber- and physical security to prevent unauthorized access and AI system exfiltration. Given the security requirements of a potentially strategically valuable dual-use technology, military-grade security, espionage protection, and red-teaming may be required. Individuals with high levels of access to AI systems should require background checks. Security auditors should assess the adequacy of security measures and compliance with information security standards through reviews and red-teaming. The security of labs should be coordinated, such that information concerning security and safety is shared between labs, either directly or through a third party such as a common government-led project or a network of security auditors. Structured access APIs [Trask et al., 2023, Shevlane, 2022] should be developed that give appropriate degrees of access to AI system developers, researchers, and auditors. Further research should be carried out on methods for verifying that code run on computing hardware is compliant with safety regulations [Shavit, 2023].

Preventing misuse of AI systems Beyond cyber- and physical security practices, AI systems introduce unique security challenges. External challenges from users include prompt injection attacks, jailbreaking, and malicious use. Mitigating these risks are active research areas. AI systems' inputs should be monitored and subject to guardrails to help tackle these risks, and auditors should be involved in assessing the extent of these risks in AI systems destined for deployment. Filtering inputs may succeed in avoiding most, but not all, jailbreaking and misuse risks. For example, if it ever became prohibited for certain systems to be placed in scaffolding that let them behave like autonomous agents (e.g. scaffolding such as LangChain or AutoGPT), then it may be possible to identify and filter the prompts used in this scaffolding, thus preventing that particular use. Auditors, in the form of internal and external red-teams, should be charged with proactively identifying vulnerabilities. Additionally, labs and governments should implement bug bounty programs, which incentivize people to find and report vulnerabilities and dangerous capabilities. Deployment audits or security audits should ensure that Know-Your-Customer (KYC) requirements and monitoring of interactions with AI systems are implemented; these may help mitigate the risks from exposure of AI systems to threat actors through prevention and monitoring.

Protection from dangerous autonomous AI systems AI systems with sufficient absolute capabilities may themselves pose security risks if acting autonomously. The affordances available to such systems should be subject to strict security assessments from external auditors. AI systems should not have unfettered freedom to acquire new affordances, such as the ability to access arbitrary code libraries or software tools. There should be clear mechanisms to recall AI systems or restrict the affordances available to them. Any interfaces that the system interacts with should be monitored for potentially dangerous use. Ongoing security assessments should be performed throughout AI system development, especially when new affordances are proposed. Scenario planning can help identify risks associated with increased affordances.

Incident response plans Institutions should have clear security protocols in place to act quickly and effectively in response to safety and security incidents. For example, at labs there should be

fail-safes, rapid response plans, and incident reporting protocols. Internal and external auditors should perform risk assessments, which should be published openly (to the extent that they do not reveal proprietary or security-compromising information). Whistleblower protections may help ensure that vulnerabilities are reported to the correct authorities; security or governance audits should ensure adequate protection. As for other dual-use technologies, emergency response plans should include government departments involved in national security to ensure that, if necessary, use of force during security and safety incidents involving highly capable frontier AI systems is legitimate.

Summary of recommendations regarding security

- Organizations with access to advanced AI systems should have military-grade information security, espionage protection, and red-teaming protocols.
- Implement strict cyber- and physical security practices to prevent unauthorized access and AI system exfiltration.
- Structured access APIs and other technical controls to enable secure development and sharing with researchers, auditors, or the broader public.
- Individuals with high levels of access to AI systems should require background checks.
- Information sharing of security and safety incidents between labs.
- The level of access to AI systems given to developers, researchers, and auditors should be appropriate and not excessive.
- Monitoring of compute usage to ensure compliance with regulations regarding the amount of compute used and how it is used.
- Prompt filtering and other input controls to prevent malicious and dangerous use, such as prohibited scaffolding methods.
- Fail-safes and rapid response plans in case the AI system does gain access to more affordances (e.g. by auto-exfiltration).
- Mandatory reporting of safety and security incidents.

2.3.7 Deployment design

Deployment of AI systems is not a binary threshold. There are many kinds of deployment, each with different consequences for risks. Deployment decisions determine who has access?; when do they get access?; and what do they have access to?

Deployment audits should assess risks from different modes of deployment for each AI system to be deployed and ensure that any regulation regarding deployment is upheld. Deployment audits aim to ensure that AI systems are not *intentionally* given excessive available affordances; by contrast, security audits aim to reduce the risk that they are given excessive available affordances *unintentionally*. Deployment audits should determine whether internal researchers vs. the general public have access to an AI system. It will also determine factors such as whether they have access to finetuning; freedom to modify the system prompt; or whether they have freedom to swap out a retrieval database for another, which could affect the absolute capabilities of the system. Another capability-relevant deployment decision is whether the system will be trained while deployed (i.e. online training), necessitating elements of both deployment audits and training design audits.

Summary of recommendations regarding deployment design

- Deployment plans should be subject to auditing.

2.3.8 Training-experiment design

A training-experiment is the procedure by which an AI system is developed. Design decisions for the training-experiment include data selection and filtering, training process; model architecture and hyperparameters; choice of deep learning framework; hardware choices; the amount of compute to use; the algorithms used; evaluation procedures; safety procedures; the affordances made available to the system during training; the properties of different phases of pretraining and fine-tuning; whether to train online or offline; etc. Given that systems trained with RL algorithms may result in

dangerous policies [Turner et al., 2019], experiments that involve RL could be subject to additional scrutiny through training design audits. The design of training experiments thus determines many of the downstream factors that are relevant to risk. Training-experiment designs should therefore themselves be subject to pretraining risk assessment, which may employ approaches such as incentive analysis [Everitt et al., 2021]. A comprehensive AI governance regime would require pre-registration, pre-approval, and monitoring of experiments involving highly absolutely capable systems trained with large amounts of effective compute and AI systems with large affordances available to them. Regulators, auditors, and AI development labs should make their own risk assessments for such experiments, with a publicly accountable institution having the final authority on whether they go ahead.

In addition to risk assessments for individual training experiments, the overall strategy into which these experiments fit (i.e. the alignment strategy) should be evaluated in ways that are accountable to the public. In order for these assessments to occur, AI system developers could be required to publish detailed alignment strategies.

Summary of recommendations regarding training-experiment design

- Require pre-registration and pre-approval of training-experiments involving highly absolutely capable AI systems trained with large amounts of effective compute and AI systems with large affordances available to them.
- Training-experiment designs should be subject to prior-to-training risk assessment.
- Require developers of frontier AI systems to publish detailed alignment strategies or to make their plans available to auditors for scrutiny.
- Require regulator approval of experiments with highly capable AI systems, large sets of available affordances, or large effective compute budgets, based on risk assessments from internal and external auditors.
- Potentially require smaller scale experiments before further scaling compute. This helps assess effective compute and predict capabilities.

2.3.9 Governance and institutions

Certain governance structures help determine the training-experiment design decisions, deployment decisions, or security decisions that are likely to be made. Auditors may be able to perform governance audits of the institutions developing general-purpose AI to ensure that incentives are aligned with safety. The AI governance and institutional landscape determines the constraints and incentives under which training experiment design decisions and security-relevant decisions are made. It is therefore important that this landscape be designed with safety in mind, rather than let it develop unconstrained under other incentives, such as profit maximization.

Mökander et al. [2023] discuss governance audits in the context of large language model development and deployment, though their work can readily be applied to broader frontier AI systems. They identify three roles for governance audits:

1. Reviewing the adequacy of organizational governance structures
2. Creating an audit trail of the frontier AI systems development process
3. Mapping roles and responsibilities within organizations that design frontier AI systems.

As the regulatory landscape is currently being designed, research should aim to anticipate and avoid potential barriers to these tasks, such as ways to ensure institutional transparency of the labs developing frontier AI systems.

Summary of recommendations regarding governance and institutions

- Labs and other relevant actors should be rendered transparent enough to regulators for effective governance.
- Regulators commission and act on governance audits when structuring the governance and institutional landscape.

3 Key areas of research

To build an auditing regime that can ensure adequate safety of AI development and deployment, further research is needed in the following key areas.

3.1 Technical AI safety research

- **Interpretability and training dynamics**: Improve methods for explaining the internal mechanisms of AI system behavior. Develop better understanding of how capabilities emerge during training through phase transitions. Use this to create predictive models that forecast emergence of new capabilities.
- **Behavioral evaluations of dangerous capabilities**: In advance of adequate interpretability methods, we should develop better behavioral methods for assessing risks from AI systems. We must develop evaluations that can serve as clear decision boundaries for particular kinds of regulatory or other risk-mitigating actions. We should improve our predictive models of how capabilities emerge in AI systems.
- **Alignment theory**: Further develop the theoretical understanding of how to create AI systems whose goals and incentives are robustly aligned with human values. This might eventually provide technically grounded metrics against which AI systems can be audited.

3.2 Technical infrastructure research

- **Structured access frameworks**: Design interfaces and structured access protocols that enable external auditors and researchers to safely analyze AI system mechanistic structure in ways that avoid misuse and proliferation.
- **Auditing of training compute usage**: Create methods to monitor and audit how computational resources are used during AI system training to ensure adherence to safety requirements and regulations. The methods would preferably be privacy-preserving, secure, and permit verification of code, data pipelines, compute usage, and other aspects of the training process.
- **Technically grounded definitions of effective compute and algorithmic progress**: A prerequisite for governance and auditing of effective compute is a technically grounded definition. Develop rigorous technical definitions and metrics for measuring the effective compute used during AI system training and for measuring algorithmic progress.

3.3 Institutional governance and policy research:

- **Accountability of auditors**: Research protocols and institutional designs that ensure accountability of auditors to the public, while also controlling potentially hazardous information flows.
- **Institutional design for transformational technology**: The current political economy of general-purpose AI development, which is currently driven by private interests, may permit less public accountability than is ideal for such a transformational technology. As the technology advances, we should consider alternative frameworks that bring AI development into a regime with a greater focus on security and public accountability, such as nationalization or internationalization.
- **Adaptive policy making and enforcement**: Build regulatory and policy-making capability to enable rapid adaptation of regulatory infrastructure as AI progresses at pace. Ensure that auditors adapt at the same pace.
- **Legal frameworks**: Explore legal tools like liability, regulations, and treaties to align AI with public interests. These provide the basis for compliance audits.
- **Frameworks for cooperation**: Develop frameworks to facilitate cooperation between governments and between companies on AI governance, even where little mutual trust exists. This may be required for auditors to operate cross-jurisdictionally, which will be required for a global approach to AI risk reduction.

Acknowledgments

Many external commenters helped greatly to improve this article, including: Markus Anderljung, Joe O'Brien, Ben Bucknall, Tom Davidson, Lisa Soder, Leonie Koessler, Jason Hausenloy, Adam Jones, Zach Stein-Perlman, Andrew Trask, Herbie Bradley, and Lennart Heim.

Contributions statement

Lee Sharkey conceived and wrote initial drafts for all sections of this article, made initial designs for the figures, and was responsible for final editing. Clíodhna Ní Ghuidhir contributed substantial editing throughout. Clíodhna Ní Ghuidhir, Dan Braun, Marius Hobbhahn provided comments and important discussion on conceptual parts of the article. Jérémy Scheurer and Lucius Bushnaq provided useful technical commentary and references. Charlotte Stix, Clíodhna Ní Ghuidhir, and Mikita Balesni provided feedback on content and assisted with figures.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

Markus Anderljung and Julian Hazell. Protecting society from ai misuse: When are restrictions on capabilities warranted?, 2023.

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

Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards publicly accountable frontier llms: Building an external scrutiny ecosystem under the aspire framework., forthcoming.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models, 2023.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, May 2022. (Accessed on 10/12/2023).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum S. Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *CoRR*, abs/1802.07228, 2018. URL `http://arxiv.org/abs/1802.07228`.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Jun Shern Chan, Michael Pieler, Jonathan Jao, Jérémy Scheurer, and Ethan Perez. Few-shot adaptation works with unpredictable data, 2022.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023.

Ege Erdil and Tamay Besiroglu. Algorithmic progress in computer vision, 2023.

Tom Everitt, Ryan Carey, Eric Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective, 2021.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022. doi: 10.1145/3531146.3533229. URL `https://doi.org/10.1145%2F3531146.3533229`.

Significant Gravitas. Significant-gravitas/autogpt: An experimental open-source attempt to make gpt-4 fully autonomous. `https://github.com/Significant-Gravitas/AutoGPT`, 2023. (Accessed on 10/13/2023).

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020.

Lennart Heim. Information security considerations for ai and the long term future. https://blog.heim.xyz/information-security-considerations-for-ai/, May 2022. (Accessed on 10/12/2023).

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023.

Langchain. Introduction | langchain. https://python.langchain.com/docs/get_started/introduction, 2023. (Accessed on 10/13/2023).

B. A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

Jieyi Long. Large language model guided tree-of-thought, 2023.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, may 2023. doi: 10.1007/s43681-023-00289-2. URL https://doi.org/10.1007%2Fs43681-023-00289-2.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022.

OpenAI. Chatgpt plugins. https://openai.com/blog/chatgpt-plugins#code-interpreter, March 2023a. (Accessed on 10/12/2023).

OpenAI. Gpt-4 technical report, 2023b.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance, 2022.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter, 2022.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression, 2023.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=1ikK0kHjvj. Featured Certification, Outstanding Certification.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm, 2021.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023.

Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. Towards best practices in agi safety and governance: A survey of expert opinion, 2023.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning, 2022.

Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring, 2023.

Toby Shevlane. Structured access: an emerging paradigm for safe ai deployment, 2022.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, 2023.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and et al. Large language models encode clinical knowledge, Jul 2023. URL https://www.nature.com/articles/s41586-023-06291-2.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative ai systems in systems and society, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng,

Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q. Tran, David R. So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, Denny Zhou, Donald Metzler, Slav Petrov, Neil Houlsby, Quoc V. Le, and Mostafa Dehghani. Transcending scaling laws with 0.1

Andrew Trask, Akshay Sukumar, Antti Kalliokoski, Bennett Farkas, Callis Ezenwaka, Carmen Popa, Curtis Mitchell, Dylan Hrebenach, George-Cristian Muraru, Ionesio Junior, Irina Bejan, Ishan Mishra, Ivoline Ngong, Jack Bandy, Jess Stahl, Julian Cardonnet, Kellye Trask, Kellye Trask, Khoa Nguyen, Kien Dang, Koen van der Veen, Kyoko Eng, Lacey Strahm, Laura Ayre, Madhava Jay, Oleksandr Lytvyn, Osam Kyemenu-Sarsah, Peter Chung, Peter Smith, Rasswanth S, Ronnie Falcon, Shubham Gupta, Stephen Gabriel, Teo Milea, Theresa Thoraldson, Thiago Porto, Tudor Cebere, Yash Gorana, and Zarreen Reza. How to audit an ai model owned by someone else (part 1). `https://blog.openmined.org/ai-audit-part-1/`, 2023. (Accessed on 10/13/2023).

Aaron D. Tucker, Markus Anderljung, and Allan Dafoe. Social and governance implications of improved data efficiency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 378–384, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375863. URL `https://doi.org/10.1145/3375627.3375863`.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power, 2019.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

Akash Wasil, Charlotte Siegmann, Carson Ezell, and Aris Richardson. Wasilezell-richardsonsiegmann+(10).pdf. `https://static1.squarespace.com/static/`

6276a63ecf564172c125f58e/t/641cbc1d84814a4d0f3e1788/1679604766050/WasilEzellRichardsonSiegmann+%2810%29.pdf, 2023. (Accessed on 10/13/2023).

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022b.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery, 2023.

Jess Whittlestone and Jack Clark. Why and how governments should monitor ai development, 2021.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation, 2022.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.