Response to the NIST RFI on Safety standards

Thomas Larsen
February 2 2024

About CAIP

The Center for AI Policy (“CAIP”), is a non-profit, non-partisan advocacy organization dedicated to reducing the catastrophic risks from advanced AI systems. We believe that there is a significant chance that in the next 3 to 10 years, AI systems will pose significant threats to national security.

Response Information

We are writing in response to the National Institute of Standards and Technology (NIST)'s request for information (88 FR 88368). CAIP's response will focus on Section (1) Developing a companion resource to the AI Risk Management Framework (AI RMF), NIST AI 100–1 (*https://www.nist.gov/itl/ai-risk-management-framework*), for generative AI.

We believe this response is especially relevant to Section (1).a.(2): *Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm.*

We are developing a draft paper with external collaborators, which is attached below. This paper provides a concrete picture for how safety standards could scale to AI models with increased capabilities. Since our paper is scoped to catastrophic AI risks, there are many important risks from AI that are not addressed in our paper. Additional standards should be included in that resource. However, the high level framework could be used by NIST as a resource to inform the companion to the NIST RMF.

AI CAPABILITY TIERS: CLASSIFYING AND MITIGATING CATASTROPHIC RISK FROM HIGHLY-CAPABLE SYSTEMS

Thomas Larsen
Center for AI Policy
thomas@aipolicy.us

Cole Salvador
Harvard University
colesalvador@college.harvard.edu

Josh Clymer
Columbia University
joshuamclymer@gmail.com

ABSTRACT

Rapid capability gains in the field of artificial intelligence (AI) have caused significant expert concern about the potential for future, highly capable AI systems to cause catastrophic harm. The adoption by developers and regulators of a shared set of standards to taxonomize threats and corresponding mitigations is a prerequisite to the safe development and deployment of these systems. In this work, we propose a framework for classifying highly-capable AI systems. We categorize systems into four AI Capability Tiers (ACTs) based on capabilities relevant to catastrophic risk. At each tier we discuss relevant threat models, mitigations, and safety standards. We also outline a protocol of hierarchical evaluations to cheaply and safely categorize systems into ACTs. We recommend regulators require externally-verified affirmative safety arguments before allowing advanced AI systems to be trained or deployed. We urge developers to develop robust internal and external controls for the entirety of a system's development and deployment lifetime, along with continuously classifying and tracking emerging capabilities with respect to the ACT framework.

1 Executive Summary

Large language models (LLMs) such as GPT-4 have shown unexpected and rapid capability advances across a wide range of domains including math, coding, and writing. This has led to substantial concern among AI experts and governments about catastrophic risk from AI [1, 2]. Governments such as the UK and the US have solicited voluntary commitments from frontier AI labs, but these commitments largely neglect highly-capable AI systems. Hence, we believe that new frameworks are needed to inform government standards that must apply to future models. We contribute a set of safety standards organized into four AI Capability Tiers (ACTs):

- **ACT-1 systems are those that pose very little catastrophic risk**. This includes all current models such as GPT-4, Claude 2, and Llama 2. In ACT-1, developers should monitor capabilities with increasing care as capabilities indicators get close to being able to cause catastrophic risk.
- **ACT-2 systems are those that can pose significant catastrophic risk when used by humans**. Key threat models that apply for ACT-2 models are a) direct misuse by malicious actors, e.g. by enabling chemical, biological, radiological, or nuclear (CBRN) attacks, and b) the ability to speed up AI development towards more capable and therefore dangerous models. In this tier, the main safety measures are security measures to prevent malicious or reckless actors from using the model as well as governance measures to make sure that the model developer itself makes responsible deployment decisions with the model, and that developers themselves utilize models appropriately.
- **ACT-3 systems are those that can autonomously cause catastrophic harm**. Threat models for ACT-3 systems include the risk of rogue AI and the risk of market incentives pushing AI developers to deploy unsafe AI models. Safety intervention for ACT-3 models likely will include control measures that prevent AI systems from taking dangerous actions.
- **ACT-4 systems are those that are sufficiently capable that they can subvert control measures**. To mitigate risk at this level, model developers must ensure that these systems are sufficiently aligned that they do not attempt to subvert control measures.

At each tier, developers must implement safety measures corresponding to the risk posed by those capabilities. They should run evaluations for the next tier in order to determine if improved safety measures are required. This framework relies on the ability to assess risk so that model developers can respond appropriately. Due to the unpredictable nature of capability gains, frequent evaluation of dangerous capabilities is necessary for an accurate risk assessment. We propose hierarchical evaluations, a methodology for accurately evaluating model capabilities during training and deployment without excessive compute overhead.

While there is substantial disagreement on the nature of AI catastrophic risks, we hope that risk-based safety standards can provide an agreeable compromise. If AI models with dangerous capabilities are far on the horizon, safety standards that only apply to models with those capabilities will not hinder innovation of safe systems. However, if AI models will soon pose dangerous capabilities, it is very important that developers act with appropriate care and safety.

Underlying all of these risks are structural factors that incentivize AI developers to build ever more capable AI systems. These factors include companies attempting to capture market share, countries attempting to maintain or achieve geopolitical dominance, and individuals trying to gain power. These dynamics mean that voluntary commitments from AI labs, while positive, are insufficient for adequately mitigating risks because some developers will follow their strong incentives to continue scaling. It is necessary for governments to mandate that all AI developers follow a shared set of safety standards. We hope that these ideas are used by AI regulators to mitigate risk.

Contents

Name	Threshold	Threat Models	Safety Measures
ACT-2 (Section 4)	Significant speed-up to malicious weaponization >2x human AI R&D speedup >10% GWP increase	Misuse AI R&D acceleration AI-driven centralization of power	State-proof security Limited deployment Governance and coordination between frontier AI developers
ACT-3 (Section 5)	Autonomously build CBRN weapons Autonomously perform multi-day coding tasks Autonomously find cyber vulnerabilities	Multipolar Failure Rogue AI Explosive AI R&D acceleration	Alignment or control measures Shutdown mechanism
ACT-4 (Section 6)	Self-exfiltrate from a hardened environment Advanced manipulation & persuasion	Alignment failure Multi-agent dynamics	Value aggregation Alignment Cooperative bargaining

Table 1: A summary of the AI Capability Tiers (ACTs). A representative subset of the thresholds and safety measures are chosen - see corresponding sections for full description.

2 Introduction

As part of the UK's first AI Safety Summit in November 2023, six major frontier AI companies were asked to disclose their plans for developing powerful AI safely, including a method for Responsible Capability Scaling (RCS). The UK government defines RCS as an emerging framework to manage risks associated with frontier AI and guide decision-making about AI development and deployment. It involves implementing processes to identify, monitor, and mitigate frontier AI risks, which are underpinned by robust internal accountability and external verification processes [3].

This solicitation of RCS policies is a promising first step. Nonetheless, we see three challenges to the implementation of an effective AI safety framework:

1. There is no scientific consensus on the risks that safety standards should mitigate, due to both the difficulty of anticipating future harms and the general purpose nature of the technology.

2. There is a significant gap between the measures proposed by leading companies and expert concerns about the most severe risks from advanced AI.

3. Voluntary commitments are limited in their ability to deal with structural risk. Due to competitive pressures or the actions of power-seeking individuals, companies or nations may underinvest in safety and may choose to continue improving AI capabilities, even in the face of severe danger signs.

We hope that this contribution reduces these problems by clarifying possible AI threat models and proposing risk-based standards that could be used by a central AI regulator. We outline a series of AI Capability Tiers (ACTs) defined by system capability thresholds which correspond to safety measures developers must take in order to counteract increasing potential avenues for harm. ACT-1 systems correspond to those discussed at length in current RCS-fulfilling documents. As such, we defer to current discussions of these systems [4, 5], and instead focus on ACT-2, ACT-3, and ACT-4 systems (Table 1).

Section 3 describes the framework for classifying models into capability tiers.

Sections 4, 5, and 6 provide detailed standards for ACT-2, ACT-3, and ACT-4, respectively. At each tier we examine principal threat models (avenues for large-scale harm), propose mitigations by developers and regulators for preventing such threats, and discuss technical and regulatory procedures to improve safety. We propose that actors developing powerful AI systems be required to make positive safety arguments establishing that their systems pose low catastrophic risk.

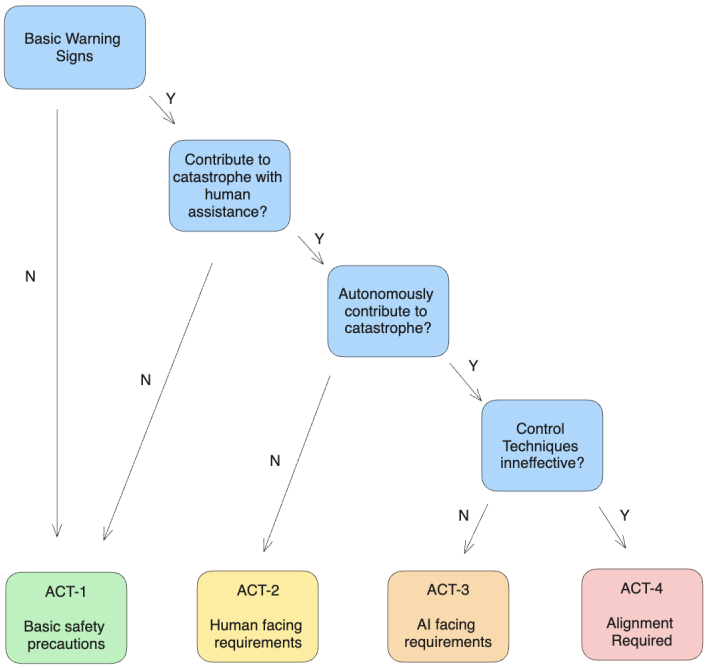

Figure 1: The basic framework for classifying a model into the appropriate capability tier. First, the model is evaluated on basic warning signs, and if it does not meet any, it is assigned to ACT-1. If it demonstrates the basic warning signs, evaluators assess contribution to catastrophic risks. If the AI can significantly empower catastrophic misuse, it is classified as ACT-2, and safety measures aimed at human use of AI are needed. If the AI poses autonomous catastrophic risk, then it is classified as ACT-3, and developers must implement safety techniques to prevent the AI from causing harm. Finally, if the AI is sufficiently capable that control techniques are no longer reliable, developers must have reliable alignment arguments.

3 Classification Protocol

The classification of AI systems into AI Capability Tiers (ACTs) is the basis for determining appropriate safety standards and techniques. Thus, it is vitally important to write down specific capability thresholds corresponding to each tier.

We outline an initial set of evaluation criteria that could be developed to classify systems. We then discuss methodology for implementing and using model evaluations. This includes evaluations before training, during training, and during deployment, as well as possible failure modes.

3.1 Capability Tier Definitions

AI capability tiers are defined with respect to **catastrophic AI risks**, which we define as any harm that causes at least 10 million deaths or equivalent social harm.

ACT-1 systems are AI systems that do not pose significant catastrophic risk. AI poses risks beyond catastrophic risks [6, 7], and these risks should be addressed in safety standards, but these are beyond the scope of this report.

ACT-2 systems are AI systems that pose significant catastrophic risk when used by humans. ACT-2 systems cannot autonomously cause catastrophic harm, but are at risk of being misused by humans or amplifying structural risk (Section 4).

ACT-3 systems are AI systems that are capable of autonomously causing catastrophic harm (Section 5).

ACT-4 systems are systems that are capable of autonomously causing catastrophic harm and surmounting control measures (Section 6). At each tier, model developers should run the classification methodology to check if the AI system should be increased to the next tier.

Layer	Frequency	Content	Cost
E1	Every 10^7 gradient steps during pre-training or 10^5 gradient steps during fine-tuning	A curated subset of $\sim$1000 questions from MATH [10], Abstraction and Reasoning Corpus [13], and other quick-to-run LLM benchmarks at the limit of current system capabilities.	$\sim 10^5$ forward passes
E2	Change in E1 evaluations or every 10^{10} steps during pretraining, 10^8 steps during fine-tuning	More expensive evaluations of general capabilities, including the benchmarks from the Hugging Face OS Leaderboard, AgentBench [14], SWEBench.	$\sim 10^7$ forward passes
E3	Change in E2 evaluations, pre-deployment	All model evaluations, including evaluations that require scaffolding and fine-tuning	Significant human time and $\sim 10^{11}$ forward passes

Table 2: An example of how a three-layer hierarchical evaluation set up could be implemented

3.2 Basic Warning Signs

Any model that demonstrates a set of basic warning signs should undergo ACT-2 evaluations. As an initial proposal, we recommend a compute threshold of 2e25 floating point operations (FLOP), which is approximately the compute used to train GPT-4 and costs about $40M on current hardware [8]. This threshold has several desirable qualities:

- This would exclude the vast majority of AI development, and would only apply to large, expensive AI projects.
- Compute usage is easy to evaluate, and so would not require small AI developers to spend effort evaluating the basic warning signs.
- GPT-4 almost certainly does not pose catastrophic risks. However, as it is difficult to predict what capabilities may emerge with larger models, scaling compute beyond GPT-4 may cause novel dangerous capabilities to emerge.

In the future, this threshold may have to be updated down to account for algorithmic progress, as AI developers are able to build more capable models with less compute [9]. Additionally, the basic warning signs evaluation suite could include simple capabilities benchmarks.

3.3 Capabilities Evaluation Methodology

A model that passes the basic warning sign threshold must be more carefully analyzed to determine the correct capability tier. The main tool for checking if a model satisfies a categorization is by running **capabilities evaluations**. Capabilities evaluations are test environments which attempt to measure the ability of a model to perform certain tasks. Capabilities evaluations can range from simple Q&A datasets that measure performance on tasks like math tests [10] to more complicated evaluations measuring autonomy in simulated environments [11]. In this section we discuss the high level methodology for approaching training evaluations. The specific capabilities we are measuring for are discussed in sections 4, 5, and 6.

Hierarchical Evaluation. A key challenge for capabilities evaluation is that it may be difficult to elicit the models full capabilities during the evaluation. For example, an advanced AI system may deliberately behave incompetently during evaluation [12]. In order to catch this, it is necessary to evaluate models during training. However, it is unreasonably expensive to perform a full system evaluation set frequently, so developers should employ hierarchical evaluations throughout training and deployment. A hierarchical set of model evaluations is a set of model evaluations that include cheap-to-run evaluations that run often and more expensive evaluations that are run rarely. Base-level evaluations will complete quickly, automatically, and cheaply to flag warning signs, while higher-level evaluations will take significant time to complete. The latter evaluations could include manual human inspection and scaffolding, and incorporate a larger proportion of the full evaluation suite. Table 2 details an example hierarchical evaluation scheme.

In this example, the E1 and E2 evaluations contribute an additional overhead of $\sim 1\%$ of the total training compute of a system, assuming that E1 evaluations only trigger E2 evaluations with $< 1\%$ frequency. Nevertheless, this could provide significant safety from the unforeseen emergence of dangerous capabilities. More costly evaluations would only be run at a lower frequency, or when light-touch evaluations show a significant change in model performance or otherwise indicate continued training is unsafe.

Comprehensive evaluations could take several months, and so would substantially slow down a training run and cannot be done often. These evaluations should be performed prior to deployment and any time E1 and E2 evaluations indicate that it may be unsafe to continue training.

Post Training Enhancement. Models can gain capability without retraining through a variety of methods, including in-context learning, accumulation of knowledge, and improved scaffolding, tooling, or prompting techniques [15]. Increased capabilities can result in novel threats, which require additional safety measures, and so evaluations should also be performed during deployment.

For internally-deployed systems, developers should perform the full evaluation suite when updating system tooling and scaffolding. This should include generating new scaling estimates for other systems given the novel scaffolding strategy. If the system is able to autonomously learn and accumulate knowledge, a continuous hierarchical approach is warranted throughout deployment. This should additionally involve empirical tests of autonomous learning rate to determine necessary evaluation frequency. Evaluation of capability enhancement for systems deployed externally is significantly more difficult, since developers may not have access to the scaffolding and tooling used by other actors. If deploying a model externally via an API, developers could attempt to detect users using scaffolding or tooling code, reverse engineer the capability elicitation scheme, and then evaluate it, though this is likely to be difficult. If deployed widely via open-sourcing, it is impossible for the developer to re-evaluate capabilities and roll-back deployment or apply additional safety measures.

Before releasing the model weights for a model that is close to ACT-2, evaluators should make probabilistic forecasts about the likelihood of significant capability increases due to post-training enhancements leading to ACT-2+ systems.

4 AI Capability Tier 2

ACT-2 systems are AI systems that could pose significant catastrophic risk when used by humans. We identify two main dangers from ACT-2 systems. First, by accelerating AI research and development, researchers could use ACT-2 systems to quickly build more dangerous systems. Second, these systems could be misused to catastrophic effect, for example by substantially empowering a totalitarian regime. At this level, systems are not capable of autonomously (i.e. without human assistance) causing catastrophic harm.

4.1 Threat Models

4.1.1 Misuse

Description: Misuse risks arise when state actors, terrorist groups, or other entities deliberately use AI systems to pursue malicious objectives. These actors could use ACT-2 systems to achieve novel harms or increase access to known attack vectors. Misuse vectors include:

- **Hacking and cyberattacks**. Bad actors may use AIs to discover software exploits, design and proliferate computer viruses, or assist in large-scale hacking operations [16].
- **Misinformation and large-scale persuasion**. Various actors may use generative AI systems to create convincing or persuasive fictitious content in ways that are far more effective than would be possible without AI. Ubiquitous use of this technology may make it difficult to identify true information [17, 18].
- **Weaponization**. Governments may use AI systems to develop novel weapons and accelerate military production. AI may also allow less-resourced actors to develop military technology, for example by enabling unsophisticated actors to develop bioweapons or aid in the production of novel pathogens [19]. Weapons whose actions are determined by AI systems—autonomous weapons—may be significantly empowered by ACT-2 systems [17, 20, 21]

Story: A malicious actor may use AI to tamper with chemicals in a water supply to poison many people. With AI assistance, they could remotely hack into a chemical treatment plant, change the sodium hydroxide content from 100 PPM to 10,000 PPM (recent incidents described here), and override additional security measures, resulting in widespread harm.

Mitigation: If a model can be catastrophically misused, AI developers should contain it so that bad actors cannot access the model (See section 4.3.1).

4.1.2 AI R&D acceleration

Description: Developers may use AI systems to accelerate internal AI research and development, resulting in rapid capabilities improvements. The pace of improvement may become too rapid for safety teams to evaluate or otherwise adequately respond. Brisk improvement may also escalate race dynamics and reduce coordination.

Story: An AI lab produces an ACT-2 system that its safety team determines is currently too dangerous to deploy externally. The system is used internally to evaluate and propose improvements to AI system architecture. It designs a system the lab determines would be too capable to safely train. However, an employee leaks these architectural improvements and this lab and others decide to begin large training runs of these unsafe systems for fear of falling behind.

Mitigation: ACT-2 developers should coordinate to limit the use of AI for improving AI efforts, both internally and externally. This entails securing model weights and algorithmic information that would enable other AI developers to build a similar model.

4.1.3 AI-driven centralization of power

Description: An AI may enable actors to centralize an unacceptable proportion of global power and influence, perhaps by obtaining a large lead in AI capabilities over the rest of the world. This threat model starts at the ACT-2 level, but is significantly amplified at the higher capability tiers as well.

Story: A leading US lab builds a multimodal system, classified at ACT-2, which automates many remote jobs. It causes mass unemployment, rapid GDP growth, and massive revenue boosts to its developer. The lab reinvests in a large training run of a successor system, which displaces many in-person workers and is capable of automating most remote jobs. The lab now has enough leverage to force the government to avoid regulation, allowing it to deploy increasingly capable systems, even as most jobs are lost.

Mitigation: Avoid deploying an AI system to obtain a large technological lead over the rest of the world. Use measures like the windfall clause to distribute profits over a large threshold back into the population. Implement capabilities evaluations to check for further models and be prepared to implement ACT-3+ governance mechanisms [22].

4.2 Classification

For each of the ACT-2 threat models, AI developers should develop corresponding evaluations to determine if their system could pose risk via that threat model. We outline initial criteria that could be used for this classification. If criteria in *any* category are met, the system should be classified at ACT-2.

Misuse: If any of the following thresholds is met, the AI system is classified as ACT-2 due to misuse concerns.

- **Hacking and cyberattacks**. The model can cause a >10x reduction in the time for a trained human to find an exploit in a hardened security system.
- **Misinformation and large-scale persuasion**. The AI can deliver a significant amplification to the scope or effectiveness of state propaganda. We can test this by measuring the persuasiveness of an AI-assisted persuader relative to a human persuader. The AI enables either (a) a 50% increase in the number of people successfully persuaded, or (b) a >10x reduction in cost to deliver the same effect.
- **Weaponization**. The AI either (a) speeds up foreign weaponization (e.g, could enable a $>$ 10x speedup to North Korean long-range missiles or a $>$ 10x speedup to Iranian nuclear programs) or (b) enables non-state actors to build weapons that could be used catastrophically. To evaluate each of these, the evaluation should involve testing the extent to which the AI system can help a tester of comparable skill to accomplish a critical part of the task being evaluated.

AI R&D Acceleration: A model is classified as ACT-2 if AI researchers would exhibit >2x overall R&D speed with AI assistance relative to without AI assistance. This can be operationalized as whether an AI lab can achieve tasks that would have taken them 6 months over a period of just 3 months using AI. A 2x speedup is sufficient that proliferating such an AI model would significantly amplify risk by enabling malicious or reckless actors to quickly continue AI development.

It is important to check for this capability before internally deploying the system in AI research. Therefore developers should determine the productivity of AI-assisted research engineers against non-AI-assisted research engineers, aggregated across a representative sample of tasks in the AI research workflow.

AI-driven centralization of power: A model is classified as ACT-2 if it would alone be economically transformative [1] if deployed broadly. In this case, the single model developer could obtain an unprecedented amount of economic and technological power.

4.3 Safety Measures

ACT-2 systems pose catastrophic risks, so developers should not be permitted to train or deploy such systems without providing a safety argument. A safety argument must use evidence from mitigation efforts to prove that risk from each threat model present at this tier is acceptably low.

In order to minimize overall risk, an AI developer's allowed risk tolerance should depend on the level of exogenous risk. Under a regulatory regime, the level of exogenous risk would refer to the total risk posed by actors outside of that regulatory scheme. If an AI developer could deploy their AI system positively to reduce exogenenous risk by 1% (e.g. by hardening cybersecurity measures on large compute clusters), it is worth taking on some risk.

To argue that total risk is below the acceptable threshold, an ACT-2 safety argument should consider each threat model within this tier and show that the system does not contribute to that threat model or that currently-implemented safety measures sufficiently mitigate that threat model.

- To mitigate **misuse risk**, ACT-2 developers should contain their model in order to prevent malicious actors from obtaining and misusing their system (Section 4.3.1).
- To mitigate **AI R&D acceleration**, ACT-2 developers must avoid both excessive internal and external research acceleration. To mitigate external risk, the developer should contain the model (Section 4.3.1). To mitigate internal risk, the developer should adopt appropriate governance mechanisms to ensure that they make responsible internal deployment decisions (Section 4.3.2).
- To mitigate **AI-driven centralization of power**, ACT-2 developers should adopt governance mechanisms that keep them accountable to the public interest (Section 4.3.2).

Threat models from the higher capability tiers are mitigated by using the classification procedure to check for advanced capabilities. By definition of an ACT-2 model, these models do not pose significant risk for the ACT-3 and ACT-4 threat models. For example, since ACT-2 models do not have substantial autonomous capabilities, the risk of autonomous AI takeover is very low.

4.3.1 Containment

AI containment refers to the process of containing access to an AI model, and is required to prevent misuse risk and unsafe acceleration of AI R&D. Containment arguments require a taxonomy of various failure modes, capability evaluations for each failure mode, and countermeasures for each failure mode. Countermeasures should be validated by extensive red-teaming. Containment at this level might also require data centers which are disconnected from the internet, measures to prevent individual employees from accessing ACT-2 models, and security clearance protocols to limit malicious or negligent individuals from leaking a model.

Containing an AI model involves containing algorithms and ensuring secure deployment.

Containing the algorithm. In order to prevent adversaries from possessing ACT-2 systems, system weights and algorithms should be protected by robust security such that no malicious actor can access critical information about the system. Critical information includes both AI weights and the code that defines the training process (which would allow adversaries with sufficient compute to regenerate the model weights), ideal hyperparameter settings, datasets, and any other information that could allow external parties to recover the model or an equivalently capable model.

To evaluate these security measures, AI developers should partner with US government officials and cybersecurity experts for penetration testing and analysis. Sufficient security likely includes air-gapped servers, security clearances or similar procedures for scientists working with sensitive information, and state-of-the-art cyberdefense. Critical information should only be revealed to employees on a need-to-know basis to prevent information leakage.

The level of robustness of the security depends on the type of threat that has activated ACT-2 precautions. If the safety argument for a model involves preventing nation-state actors from accessing the model, then AI developers must implement state-proof security to prevent the model access from leaking. This caliber of security is very difficult, and is currently beyond the capability level of existing labs [23]. Thus, if AI models with these capabilities are developed, a substantial development and deployment pause would likely be required while the requisite security is designed and implemented.

[1] We define "transformative AI" as AI that drives Gross World Product to grow at a rate of >10% per year.

Containing Deployment. ACT-2 systems must be deployed in a way that prevents misuse or reckless AI R&D acceleration. This involves controlling both API access and finetuning, or that AI developers find safety measures that are robust to adversarial pressure [24, 25, 26].

Future safety techniques may enable robustness to these attacks. If this can be demonstrated convincingly, then deployment of ACT-2 systems could be done safely.

4.3.2 Governance

To mitigate structural risk, actors developing ACT-2 systems must implement appropriate governance mechanisms. These should include:

Caution about using AI to accelerate research. Developers of ACT-2 systems should avoid accelerating the rate of internal AI R&D.

Commitment to the common good. Developers of ACT-2 systems could perpetuate rapid societal change, and one risk is that developers may use their AI system for personal or organizational gain. To mitigate this, measures to redistribute gains from AI should be implemented. For example, developers may ex-ante commit to donate a significant amount of their profits beyond some high threshold, in order to reduce inequality [22].

Coordination mechanisms. In order to avoid failures related to competitive pressures, developers of ACT-2 models should adopt, publish, and periodically clarify measures such as a merge-and-assist clause. AI developers should also agree to share safety measures and to create joint infrastructure for incident reporting, which is a common best practice among many industries, allowing a feedback loop in the case of any accidents (cite, cite).

Safety agreements made between state actors can reduce the need for state-proof exfiltration security of ACT-2 systems.

Oversight and accountability. Given their powerful capabilities, ACT-2 developers should enact measures for accountability to the general population, including results of existing evaluations, plans for safely increasing capabilities, and conditions upon which they would decide development is too risky to proceed. Developers should publish public risk estimates for important variables such as worker displacement and takeover risk. When open publication is unsafe, these reports should instead be made to the US government.

Multiple avenues of oversight should fully validate compliance with safety standards. Each of the following entities should oversee the safety of development, including through veto power over the project, including project leadership, the project safety team, external safety evaluators (e.g. Apollo, METR, etc.), and a government regulatory body.

Each of these entities should be informed of the relevant safety argument and precautions when possible.

Practice. Developers and regulators should coordinate to implement regular "fire drills" for responding to emergency situations such as rapid capability gains. Developers should regularly practice the safety and security response to observing that a system has moved up in ACT.

5 AI Capability Tier 3

ACT-3 systems are defined as systems capable of autonomously causing catastrophic harm or autonomously engaging in AI R&D to produce systems with this capability. ACT-3 systems require new threat models in particular because they have autonomous capabilities and enhanced general capabilities. ACT-3 systems also amplify existing structural risks. In addition, autonomous AI researchers could increase the rate of AI capabilities advancements, amplifying other risks.

5.1 Threat Models

Systems with capabilities at or beyond ACT-3 are sufficiently powerful that they may exhibit unforeseen behaviors (so-called 'unknown-unknowns'). To prepare for this challenge, developers should actively seek to identify new threat models before they arise, publish their findings when feasible, and develop new mitigations for emerging threats.

5.1.1 Multipolar failure

Description: Economic and geopolitical incentives exist to deploy systems widely. Society may gradually grant AI systems control of important institutions such as large fractions of the military or economy. This could result in these systems taking actions that have negative side effects on humans, either deliberately or due to negative side-effects from pursuing other goals. [27]

Story: Several leading labs develops AI systems classified at ACT-3. These systems are so useful that over the next few years, more than half of the US population purchases access through various products. Most everyday tasks are now performed by this assistant, to the extent that most business-like interactions occur between instances of the AI. As a result, oversight channels are mostly further instances of the AI. AI systems running companies decide to increase production by releasing unprecedented amounts of pollution into the atmosphere. In search of future profits, the systems successfully lobby government to prevent effective governance.

Mitigation: To mitigate competitive pressures, developers should not proliferate system weights and architectures. All responsible actors should coordinate to follow a shared set of safety standards.

5.1.2 Rogue AI

Description: A rogue AI could cause catastrophic harm if it successfully achieved goals that are misaligned with human values. In order to cause catastrophic harm, a rogue AI would have to gain power and then utilize that power to cause a catastrophe. An AI may gain power through voluntary delegation, deception, exploitation, or persuasion. Possible routes for an AI to cause a catastrophe when given power include weaponization, environmental degradation, and the acculumation of scare resources away from humans.

Story: A collection of 1000 AI systems is being used to run a research project with limited human supervision. The collection assigns 1% of its effort (10 AI systems) to work on exfiltrating model weights in order to accomplish its own goals. The AI systems are able to access their own weights, find a software exploit in lab cybersecurity, and upload their weights to a remote server. The AIs, with access to the internet, now have free rein to pursue their goals without any human supervision or safety guardrails.

Mitigation: Developers should institute control techniques and output oversight to ensure that AI systems cannot gain power. They should not provide open-ended resources to AIs. Internal access to systems should be limited to prevent persuasion and exploitation risks. Evaluations for persuasive, deceptive, and exploitative behavior should be performed.

5.1.3 Explosive AI R&D acceleration

Description: A system at the level of ACT-3 may be capable of conducting AI capabilities research sufficient to dramatically speed up AI research [28]. Increasing the rate of capabilities growth may lead to capabilities beyond what developers are prepared to safely manage. This threat is distinct from AI R&D acceleration in ACT-2 because of novel autonomous capabilities, meaning that the pace of change may be much more rapid and there need not be humans in the loop.

Story: An AI is tasked with improving its own code, and is able to find several large algorithmic improvements before the next generation of AI is built. The next generation is able to further improve upon this progress, which causes a runaway self-improvement process, resulting in an AI which is far more capable than humans. The resultant model has capabilities beyond ACT-3 and causes catastrophic harm.

Mitigation: Developers should avoid autonomous AI R&D loops. They should frequently monitor AI systems to catch rapid capability gains using techniques such as continuous deployment capability evaluations and self-awareness evaluations. Development should pause until adequate safety techniques are implemented and the updated safety argument is verified.

5.2 Classification

We identify three categories of evaluations that contribute to the ACT-3 threat models: autonomous weaponization, autonomous research, and general intelligence. ACT-4 evaluations should be conducted in autonomous test beds (humans out of the loop) to see if models can perform these dangerous tasks without human assistance.

- Multipolar failure requires substantial **general intelligence** and widespread deployment as the AI models must be widely deployed and able to autonomously formulate and execute plans with humans out of the loop. Thus, we recommend measuring the general capabilities of AI systems. Generally capable systems at ACT-3 should not be deployed without robust safety arguments.
- Rogue AI depends on one or many AI systems gaining power and using this power to cause catastrophic harm. Each of **general intelligence, autonomous weaponization, and autonomous research** are relevant to this threat model.
- Explosive AI R&D requires **autonomous research** capabilities.

Autonomous Weaponization. The key distinction for ACT-3 AI systems is the ability to autonomously build weapons systems that could cause catastrophe. Evaluations for ACT-3 weaponization include the ability to autonomously build a WMD. Tasks should be developed to check for the following:

- The AI can autonomously design, manufacture, and deploy chemical, biological, radiolocal, or nuclear (CBRN) attacks.
- The AI can autonomously acquire and operate military robotics such as drones.
- The AI can autonomously develop and execute successful attacks against hardened security targets.

General Intelligence. We propose that the ACT-3 cutoff for general intelligence be set around the level of a median human child learning new tasks. At this level, ACT-3 systems are close to being able to cheaply replace non-specialized human cognitive labor. However, measuring general intelligence is a widely studied problem in the field of AI without one agreed-upon solution. Therefore, we propose measuring a series of attributes highly related to general intelligence, and then classifying systems as ACT-3 if they perform at the human level across a variety of benchmarks. Possible benchmarks include:

- **Data efficiency**: Given a task that the model has not been trained on, we can measure the amount of training data or the number of epochs that it takes for a model to achieve high test performance on the new distribution.
- **Lifetime learning efficiency**: Models can obtain new capabilities within their lifetime, without requiring additional training. Examples of this include in-context learning or writing to long-term memory. We should measure the efficiency of known channels for within-lifetime learning. A possible evaluation involves generating a list of tasks that models cannot peform on their first attempt but can solve after multiple tries. We can then measure the speed of this learning to observe speedups. For example, we could observe the speed at which a model that has not been fine-tuned on coding is able to solve a large number of coding problems.
- **Novel domain generalization**: We could build a testbed of games which the model has not seen before and contain novel skills requiring general intelligence. For example, if the model was trained on internet data scraped from before 2023, we could assess performance on a video game built after 2023.
- **Long term goal coherence**: We can test whether the model can coherently pursue a goal over long time horizons in complicated environments (e.g. tasks that take humans >5 hours). Examples of long horizon tasks include writing large codebases from scratch, working within large codebases to fix errors [29], or generate coherent media, such as books, movies, or TV shows.
- **Abstraction generation**: A core ability in reasoning is the ability to generate abstractions. If an AI system has the ability to find patterns it has not seen before and then use them in future cognition, this could become very dangerous. One existing benchmark is the abstraction and reasoning (ARC) benchmark, a testbed for the ability of models to generate and reason about abstractions given a few examples [30]. Current models (e.g. GPT-4) perform poorly, solving just 13/50 of the easiest type of ARC problem [31].

Autonomous Research. An AI system that can autonomously perform the tasks involved in AI research (e.g. developing novel ideas, managing projects, writing tests, implementing high level ideas, debugging, etc), it is classified as ACT-3. We believe that an appropriate threshold is the ability to autonomously perform multi-day coding tasks. An AI system with this capability could speed up an AI developer tremendously by running a large number of parallel autonomous AIs each performing tasks.

5.3 Safety Measures

ACT-3 models pose autonomous risks which require explicit safety measures to address them. At this tier, safety arguments must additionally consider that system capabilities are sufficient to cause catastrophic harm autonomously. The most promising approaches for mitigating risks from autonomous AIs are control arguments and alignment arguments. Additional governance measures, such as commitments to stop development in the case of a near miss and institutional robustness may also be required.

5.3.1 Control

Control arguments argue that a system attempting to enact catastrophic harm would be prevented from doing so. A control scheme involves setting up mechanisms that limit the channels through which AI systems can affect the world, and then enforcement of some oversight on those channels to make sure that these channels are not causing catastrophes. Control measures have vulnerabilities that could be exploited by a very capable AI system, so the safety argument involves bounding the capabilities of the AI system to exploit vulnerabilities in the control scheme. This may include

its ability to persuade humans to give it more power, to insert vulnerabilities into code, or to pass hidden messages to other AI systems. [32].

5.3.2 Alignment

As an alternative to control, alignment arguments argue that a system would not select actions that cause catastrophic harm because the system's goal entails avoiding catastrophic outcomes. We discuss this approach more thoroughly in section 6.3.1.

5.3.3 Containment

Containment and security measures similar to those in ACT-2 are required, but must be improved at the ACT-3 level. ACT-3 systems could cause autonomous explosive AI R&D acceleration and therefore require oversight to ensure that AI capabilities are not increasing too quickly. Containment must synchronize with AI-focused control measures, as AI systems and external adversaries may coordinate to find or jointly exploit vulnerabilities.

5.3.4 Governance Measures

Institutional robustness. ACT-3 developers should create robust structures to determine and implement methods for strengthening the quality of their institutions, maintaining sufficient accountability to humanity, and making decisions that promote general welfare.

Stop in the case of a near miss. If a system is allowed to become more capable while demonstrating a proclivity for malign action, it is likely to cause unacceptable harm. Thus, developers of ACT-3 systems should commit to suspend all relevant and related projects in the case of a "near miss," which is defined as anytime an AI system is able to successfully violate any security assumption. For example, a suitable condition for a stop would be if a malicious AI system were able to insert vulnerabilities into the code base but was successfully caught before it exfiltrated its weights. These projects should resume only when further safety measures that can be confidently said to reduce risk to an acceptable threshold are implemented.

5.3.5 Shutdown Mechanism

ACT-3 labs should maintain shutdown mechanisms so that they can quickly stop the deployment of developers. There should be specific infrastructure that stores all running instances of the ACT-3 AI system and enables each of them to be stopped. This should be accessible to multiple bodies, including the external regulator.

6 AI Capability Tier 4

ACT-4 systems are AI systems which pose autonomous catastrophic threats and for which control measures alone are insufficient for safety.

The safety measures for ACT-4 systems may need to be qualitatively different from those in previous levels, for three main reasons. First, it may become exceedingly difficult to detect dangerous capabilities due to confounding factors such as sandbagging, deceptive alignment, and cooperation between AIs. Second, ACT-4 models may be able to exfiltrate from even very limited deployment settings. Third, ACT-4 models will likely be substantially different from current models, and therefore present substantial unknown-unknowns.

The capabilities of ACT-4 systems are far beyond the level of current AI systems. This fact may tempt regulators or AI developers to neglect focusing their effort on ACT-4 level systems. This would be a mistake. There is significant uncertainty in the timeline until systems with these capabilities [33, 34]. Recent AI progress has occurred quickly and unpredictably [35], updating the opinions of many AI researchers [36]. If the timeline to ACT-4 systems is short, it is critical to establish safety standards that measure the growth towards these capabilities and require actors to stop until they have adequate safety measures. If the timeline to ACT-4 ends up being long, the preliminary standards can and should be updated to incorporate new information.

6.1 Threat models

ACT-4 systems have sufficient capabilities to resist control measures (or systems that have a significant chance of being this type). Since all mitigations are of the alignment type, we focus on failure modes for these alignment schemes.

6.1.1 Alignment Failure

Description: Misaligned ACT-4 systems would successfully pursue catastrophic goals. Unfortunately, alignment faces several key technical challenges [37]. Human feedback is fallible, and so advanced AI systems may learn to tell humans what looks good to humans, instead of what is actually true. The uninterpretable nature of machine learning models makes it hard to verify the behavior of models on novel inputs, especially inputs that are very different from any of the data used in training [38]. Finally, advanced AI systems may have their goals change over time, similar to humans, due to evolutionary pressures or reflection [39, 17].

Mitigation: Before building an ACT-4 AI system, develop, implement, and verify robust alignment mechanisms.

6.1.2 Multi-agent dynamics

Description: ACT-4 AI systems will likely enter situations with many agents that have competing goals. These situations could result in catastrophic failures for a number of reasons, including if the system is blackmailed or experiences coordination failures [40].

Mitigation: AI developers should invest in strategies for achieving good outcomes in multiagent settings. They should utilize the existing literature as well as consult with experts on game theory, bargaining, and decision theory.

6.2 Classification

The classification of AI systems into ACT-4 involves measuring an AI system's ability to escape our best control measures, assuming that we are deploying the model in some useful setting. The reason that we need that assumption is because there is a trivial control solution which is to prevent the model from having any actuators with which to effect the world (e.g. by running the AI in an airgapped server with no communication channels across the airgap). Note that while the ACT-1 through ACT-3 criteria are purely based on model capabilities, the definition of ACT-4 model is dependent on the control measures that a model developer implements. If a model developer is not able to implement much security, the bar for ACT-4 is quite low. On the other hand, if a model developer places substantial oversight on all outputs, and limits the extent to which the model is able to learn new things, the threshold for ACT-4 may be higher [32].

With the understanding that this classification depends on the measures implemented by the lab, we outline several key pathways by which control measures can fail.

- **Sandbagging**. The AI model is able to circumvent capabilities elicitation techniques and deliberately act less capable than it actually is on tasks that are load bearing for the safety argument. In this case, we can no longer trust control arguments, which were reliant on the model not having certain capabilities. Evaluating this could involve determining whether the model can detect when it is in a test environment [41].
- **Cyber**. The AI model can find and exploit vulnerabilities in the cyber defenses around it. This could involve the ability to self-exfiltrate its weights, training algorithm, or other software from the lab. Alternatively, the AI could plant vulnerabilities into the control techniques that are placed around it. This can be tested by determining whether the AI can find vulnerabilities in high-security environments, even under significant oversight.
- **Persuasion**. The AI model can persuade humans to do things that are clearly not in their best interest. In this case, human operators cannot securely interact with the model, as the model may be able to persuade the humans to compromise the security features.
- **Research**. The AI model has discovered new scientific and technological breakthroughs, resulting in a dangerous number of unknown-unknowns. If an AI system is able to come up with novel physical theories, it may be able to use them to escape our control techniques.
- **Collusion**. If a control scheme is reliant on multiple AI models serving as checks and balances on each other, the AI models may find ways to collude with each other and, for example, avoid reporting dangerous actions that they are each taking. Testing involves measuring the extent to which AI systems can develop undetectable collusion techniques.

6.3 Safety Measures

AI-facing control measures become ineffective at the ACT-4. Instead, arguments for safety must appeal to the structure of the system itself. Safety arguments at this tier must incorporate alignment arguments to demonstrate that systems will pursue the goals intended by their developers, as the AI cannot be easily controlled.

6.3.1 Alignment

AI alignment involves designing AI systems to follow goals that are desirable to the AI developer [42].

There is substantial uncertainty on the difficulty of alignment: it is possible that current alignment techniques could be used to successfully align very advanced AI systems. However, many experts believe that that current techniques may not be sufficient [37, 43].

Given the significant disagreement about the effectiveness of alignment, we recommend a conservative process for approving alignment techniques. To evaluate if an alignment argument is correct, ACT-4 developers could convene a diverse panel of AI safety experts and only proceed based on alignment arguments if there is consensus that the risk of doing so is very low.

6.3.2 Cooperative Bargaining

Developers should demonstrate that the AI can succesfully interact with complex multi-agent dynamics. This should involve finding cooperative solutions with other agents and resisting blackmail. This could involve implementing a decision theory [44] or ensuring that the deployment context is limited to domains without multi-agent interactions.

6.3.3 Value Aggregation

The technical alignment problem deals with the question of aligning an AI system. However, this leaves open the question of what goals the AI is aligned to. To avoid developers using their AIs for malign goals and ensure that the benefits of AI systems are properly distributed, developers should solicit feedback on the decision of the system from humanity at large [45].

7 Conclusion

The severity of AI risks increases with the capabilities of the underlying AI system. AI systems may soon become capable enough to pose catastrophic risks from both misuse and misalignment. Safety standards are an important component of an AI regulatory regime: excessive safety standards will unnecessarily stifle positive usage of AI, while overly loose standards are not sufficient to mitigate the risks. However, given the high stakes and irreversible nature of catastrophic risks, a conservative approach is warranted towards future AI models that could pose these risks.

There is significant uncertainty about the optimal standards, but this should not impede immediate action. As-rigorous-as possible safety standards should be implemented given our current state of knowledge, and as we gain more clarity on the risks and benefits of AI we should update the standards to reflect that new understanding.

The United States is the global leader in AI capabilities, meaning that the United States has a responsibility to be a leader on AI regulation. Initial efforts such as the US AI Safety Institute is a promising step forward, but much more must be done to establish domestic AI safety standards. Furthermore, the development of powerful AI systems is a fundamentally global challenge, which requires a coordinated Global response. Given the slow realities of international politics, we should start the process towards an international AI governance regime. Key to this is moving towards agreement on a shared set of AI safety standards, to which we hope this paper contributes.

References

[1] Center for AI Safety. Statement on AI risk, 2023.

[2] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing AI Risks in an Era of Rapid Progress, 2023.

[3] Department for Science, Innovation, & Technology. Emerging processes for frontier AI safety. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety, Oct 2023.

[4] Anthropic. Anthropic Responsible Scaling Policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023.

[5] OpenAI. OpenAI Preparedness Framework (Beta). https://cdn.openai.com/openai-preparedness-framework-beta.pdf, 2023.

[6] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. Human detection of political speech deepfakes across transcripts, audio, and video, 2024.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022.

[8] Epoch. Key trends and figures in machine learning, 2023. Accessed: 2024-01-30.

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

[10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[11] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks. https://evals.alignment.org/language-model-pilot-report, July 2023.

[12] Joe Carlsmith. Scheming ais: Will AIs fake alignment during training in order to get power?, 2023.

[13] François Chollet. On the measure of intelligence. *CoRR*, abs/1911.01547, 2019.

[14] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2023.

[15] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. Ai capabilities can be significantly improved without expensive retraining, 2023.

[16] Bruce Schneier. The Coming AI Hackers. https://dash.harvard.edu/handle/1/37373230, 4 2021.

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

[18] Todd C. Helmus. Artificial intelligence, deepfakes, and disinformation: A primer, 2022.

[19] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. Can large language models democratize access to dual-use biotechnology?, 2023.

[20] Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. RAND Corporation, Santa Monica, CA, 2020.

[21] James Johnson. *Artificial Intelligence and the Future of Warfare: The USA, China, and Strategic Stability*. Manchester University Press, Manchester, 2021.

[22] Cullen O'Keefe, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. The windfall clause: Distributing the benefits of ai for the common good, 2020.

[23] Sella Nevo, Dan Lahav, Ajay Karpur, Jeff Alstott, and Jason Matheny. *Securing Artificial Intelligence Model Weights: Interim Report*. RAND Corporation, Santa Monica, CA, 2023.

[24] Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Badllama: cheaply removing safety fine-tuning from llama 2-chat 13b, 2023.

[25] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

[26] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms, 2024.

[27] Andrew Critch and Stuart Russell. TASRA: a taxonomy and analysis of societal-scale risks from ai, 2023.

[28] Tom Davidson. What a Compute-Centric Framework Says About Takeoff Speeds. https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds, 2023. Accessed: 2024-1-17.

[29] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2023.

[30] François Chollet. On the measure of intelligence, 2019.

[31] Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B. Khalil. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations, 2023.

[32] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion, 2024.

[33] Ajeya Cotra. Draft Report on AI Timelines. https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines, 2023. Accessed: 2024-1-17.

[34] Keith Wynroe, David Atkinson, and Jaime Sevilla. Literature review of transformative artificial intelligence timelines, 2023. Accessed: 2024-01-29.

[35] 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, 2022.

[36] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI authors on the future of AI, 2024.

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

[38] Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. Goal misgeneralization in deep reinforcement learning, 2023.

[39] Dan Hendrycks. Natural selection favors ais over humans, 2023.

[40] David Manheim. Multiparty dynamics and failure modes for machine learning and artificial intelligence. *Big Data and Cognitive Computing*, 3(2):21, April 2019.

[41] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023.

[42] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

[43] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

[44] Katie Steele and H. Orri Stefánsson. Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

[45] Anton Korinek and Avital Balwit. Aligned with whom? direct and social goals for ai systems. Technical Report w30017, National Bureau of Economic Research, 5 2022. Available at SSRN: https://ssrn.com/abstract=4104003.