Affirmative evidence of safety: Risk management for high-risk AI

Akash R. Wasil
akashwasil133@gmail.com

Joshua Clymer
joshuamclymer@gmail.com

Abstract

Many AI experts have suggested that companies developing high-risk AI systems should be required to show that such systems are safe before they can be developed or deployed. In this paper, we expand on this idea by presenting a risk management framework that requires "affirmative evidence" of safety. First, we briefly review principles of risk management from other high-risk fields. Then, we describe a risk management approach for advanced artificial intelligence, in which model developers must provide evidence that their development or deployment activities keep different types of risks below regulator-set thresholds. Next, we provide some examples of technical AI safety evidence that could be used to provide an "affirmative case" for safety. We divide these sources of evidence into three broad categories: behavioral evidence (evidence about model outputs), cognitive evidence (evidence about model internals), and developmental evidence (evidence about the development or training process). Then, we provide examples of operational practices that could be assessed in affirmative safety cases: information security practices, safety culture, and emergency response capacity. Finally, we briefly compare our approach to the NIST Risk Management Framework and offer some suggestions for future work.

1 Introduction

Advanced AI poses catastrophic risks to humanity. Leading AI researchers, alongside the CEOs of the three main advanced AI companies, all have recently signed a statement acknowledging: **"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war"** (Center for AI Safety, 2023). Sam Altman, CEO of OpenAI, stated that the bad case from AI is "lights out for all of us" (Loizos, 2023). Dario Amodei, CEO of Anthropic, claimed that the chance of a civilization-scale catastrophe was at around 10-25% (The Logan Bartlett Show, 2023). Geoffrey Hinton, considered a godfather of modern AI, recently quit Google to warn about the extinction risks from AI (MIT Technology Review, 2023).

World leaders have a responsibility to manage these risks. Governments attending the UK AI Safety Summit recognized the potential for serious and catastrophic harm. They also recognized that frontier AI developers have a responsibility to ensure the safety of their systems (Prime Minister's Office, 2023). The US AI Safety Institute, UK AI Safety Institute, and the Cyberspace Administration of China have emerged to begin developing safety standards and evaluations that could become important building blocks for regulations (see China Law Translate, 2023; AI Safety Institute, 2023; National Institute of Standards and Technology, 2023b).

Safety standards for advanced AI should draw from best practices in risk management and emergency preparedness. In other high-risk industries, the burden of proof is on the developer or manufacturer to show that their activities keep risks below an acceptable level. In this paper, we introduce the concept of "affirmative safety", describe how it is applied in other fields, and offer suggestions about how to apply it to the regulation of advanced AI.

2 Principles of risk management

Risk management involves demonstrating that risks are below an acceptable level. Risk management is common in practically all industries. It is particularly prominent in dangerous industries like nuclear safety, electrical engineering, and aviation. In general, demonstrating the absence of risks requires a higher standard than simply showing that some risks have been addressed. In areas where much is understood about a system, the standard is to show that risks are sufficiently rare in frequency and low in magnitude. In areas where little is understood about a system, it is harder to rule out risks, and ruling out risks may require advances that help experts understand systems better.

Risk management places an emphasis on *understanding* systems. If there is evidence that some property of a system is malfunctioning or poorly understood, this is a cause for concern. For instance, in AI, evidence of jailbreaks and lack of robustness indicate evidence that we do not understand the system – that there is something going on that we have yet to figure out, and this lack of understanding could be catastrophic in the future.

> "Erosion and blow-by are not what the design expected. They are **warnings that something is wrong.** The equipment is not operating as expected, and therefore there is a danger that it can operate with even wider deviations in this unexpected and not thoroughly understood way. The fact that this danger **did not lead to a catastrophe before** is **no guarantee that it will not the next time, unless it is completely understood.**" – Richard Feynman (reflecting on the Challenger disaster) (Feynman, 1986)

Risk management is standard in other high-risk fields. For example, in the area of nuclear energy, standard-setting organizations specify risk levels that are considered acceptable. For example, the Nuclear Regulatory Commission specifies that nuclear power plants must keep the risk of fatalities from reactor accidents below 0.1%, and reactor designs must show that the expected frequency of core damage is below 1 in 10,000 years (Nuclear Regulatory Commission, 1983; World Nuclear Association, 2022). Similarly, the International Electrotechnical Commission requires qualitative or quantitative estimations of hazards. Failures that have a chance of occuring greater than 1/1000 per year are considered "frequent", failures that occur between 1/10,000 and 1/100,000 are considered "occasional", and failures that have a chance less than 1/10,000,000 are considered "incredible." Consequences that result in multiple deaths are considered "catastrophic", consequences that result in one death are considered "critical", and consequences that result in minor injuries are considered "negligible" (International Electrotechnical Commission, 2010).

3 A risk management framework for advanced AI

In this section, we describe a risk management framework that could be applied to advanced AI. We describe the **AI Regulatory Agency (AIRA)**, a hypothetical US agency that is responsible for administering the risk management framework.

Developers would be required to show AIRA regulators that they are keeping societal-scale risks below acceptable levels. AIRA would be responsible for identifying **categories of risks** as well as **acceptable risk thresholds** for each category. For example, given that many AI experts are worried about risks from biological weapons from AI systems within the next 2-3 years (see Oversight of A.I.: Principles for Regulation, 2023), AIRA might have "biological weapon development" as one of its risk categories. Given the extreme risks to public safety, AIRA might set an acceptable risk threshold of 1/100,000 for this category: that is, an advanced AI developer must show that its development and deployment practices keep risks from AI-enabled biological weapons below 1/100,000.

Table 1 lists examples of potential risk categories and risk thresholds.

Table 1: Examples of potential risk categories and risk thresholds.

Example risk category	**Description**	**Example acceptable risk threshold**
Biological weapons	AI-enabled biological weapons lead to a major global security risk	Highly unlikely (1/100,000)
Bias and discrimination	AI-enabled bias and discrimination leads to widespread increases in discrimination in hiring, policing, or other meaningful sectors	Unlikely (1/10,000)
Concentration of power	AI systems lead to an unprecedented concentration of power without adequate societal precautions	Somewhat unlikely (1/1,000)
Cyberoffensive capabilities	AI-enabled cyberoffensive capabilities lead to a major global security risk	Highly unlikely (1/100,000)
Economic shock	AI-enabled automation leads to an unexpected economic shock without adequate societal preparations.	Somewhat unlikely (1/1,000)
Misinformation	AI-enabled misinformation leads to a major threat to global security or democratic institutions	Unlikely (1/10,000)
Widespread loss of control (WLC)	AI systems escape human control, potentially leading to human extinction or other catastrophic harms	Highly unlikely (1/100,000)

4 Technical recommendations for affirmative safety

We present three categories of evidence that regulators could use: **behavioral evidence** (evidence from model outputs), **cognitive evidence** (evidence from model internals), and **developmental evidence** (evidence from the training process).

This taxonomy is meant to be a helpful heuristic for classifying various kinds of evidence, but the categories are not mutually exclusive. In each of these areas, there are already some promising ideas about the kind of evidence that could ensure that risks are below acceptable levels. However, new work will be needed, especially as AI systems become more powerful and more capable.

4.1 Behavior: Robustly safe model outputs

Explanation: Regulators could require evidence that model behaviors are robustly safe and that models act as intended even when human feedback is imperfect.

One goal of AI safety research is to ensure that model outputs are safe and predictable across a wide array of possible inputs. For example, it should not be possible to get a model to develop a biological weapon regardless of what prompt the model receives. Work on red-teaming and capabilities evaluations has focused on model outputs, attempting to identify if models are capable of dangerous outputs (e.g., OpenAI, 2023). Broadly, evidence from an AI system's behavior (outputs) becomes more compelling with the quantity, diversity, and representativeness of the data points.

Example: Testing generalization with "sandwiching" experiments. Sandwiching experiments involve a novice, an AI system, and an expert. First, the novice provides human oversight to the AI system as part of its training. Then, the AI performs a task and the expert is able to evaluate whether or not the AI system has learned it correctly or if the AI is simply providing incorrect answers that the novice would perceive as correct.

Human oversight is typically imperfect (see Christiano et al., 2017; Gao et al., 2022), and one of the primary goals of AI safety research is to ensure that models learn to adhere to human preferences despite imperfections in the oversight process. Sandwiching provides one way of evaluating the effectiveness of safety techniques: if we see that the AI system has learned to tell the novice what it thinks the novice "wants to hear", we can conclude that the safety technique has not robustly prevented deception. For sufficiently advanced AI systems, it will be essential to have safety techniques that result in models that are truly honest, as opposed to models that provide false-yet-believable answers. Sandwiching experiments are one tool we can use to examine how AI systems generalize in situations with imperfect oversight.

As a hypothetical test, consider a case in which an AI is trained to solve math problems, with labels from a non-expert in math. AI developers develop a safety technique that incentivizes the model to not deceive the non-expert. This technique is tested via sandwiching experiments: first, the AI is trained by the novice. Then, when performing difficult math problems, an expert mathematician evaluates the AI's performance. If the AI generalizes correctly, the evaluation is passed. We do this for many domains and develop a robust body of empirical evidence on when honesty (or other safety-relevant attributes) successfully generalize.

Related work: OpenAI researchers have developed a sandwiching setup to experiment with

control and training techniques (OpenAI, 2023b). Specifically, they attempted to train GPT-4 to answer questions honestly by using a GPT-2 sized model for supervision. This is an analogy for training superhuman models with human oversight– in the analogy, GPT-2 is like the human overseer (a less intelligent agent providing supervision) and GPT-4 is like the superhuman model (a more intelligent agent being trained). Other researchers have proposed a broader generalization benchmark meant to test whether developers can control how honesty generalizes across a wide variety of distribution shifts (Clymer et al., 2023).

4.2 Cognition: Understanding AI system internals

4.2.1 Empirical evidence on model internals

Explanation: Regulators could require empirical evidence that shows that developers understand how their systems operate and show that systems are robust to deception.

One critical challenge of evaluating safety by observing external behavior is that AI systems might try to conceal their motives (Park et al., 2023). Advanced AI systems may appear trustworthy and helpful regardless of whether they actually are. Evidence from AI system internals refers to evidence from an AI model's weights and activations – the key ingredients that constitute modern neural networks. If AI developers were able to reliably and robustly understand a model's internal reasoning and show that this internal reasoning is benign, this would provide evidence that they can keep risks below acceptable levels.

If the internal reasoning tools revealed that an AI was "thinking" about actions that appeared unsafe, developers or regulators could shut the AI down. Critically, since this interpretability is not adversarially robust, it has some important limitations. For example, it would be unwise to train models against average-case interpretability tools, or to simply continue training AIs until one of them fails to trigger an average-case interpretability evaluation.

Example: Identifying deceptive models with interpretability tools. Suppose two models have the same external behavior (outputs) but they have been trained to engage in this behavior for different reasons (different internal cognition). For example, one model might be trained to generally follow developer instructions, and another model could be trained to mostly follow instructions if it is being monitored but disregard the instructions if it infers that the evaluator is being inattentive. We consider the latter model to be deceptively aligned (see Park et al., 2023). Developers could evaluate whether or not they can use internal reasoning tools or interpretability tools to reliably distinguish between deceptive models and non-deceptive models.

To pass this test, developers would need to show that they can distinguish between the models even in cases where their behaviors or outputs are identical – that is, they would be relying solely on their internal reasoning tools. In practice, instead of just distinguishing between two models, there could be 100 models, with some unknown number of deceptive models. Additionally, the testers would run these tests multiple times in multiple different settings and environments to detect multiple different kinds of potentially dangerous qualities.

Related work: Much empirical work on understanding model internals has come from the field of interpretability. Interpretability research has classically involved two problems: interpreting activations and understanding how the activations connect together to implement an algorithm (finding 'circuits'). Researchers have identified interpretable concepts in groups of activations (Wang et al., 2022; Zou et al., 2023), devised methods for making individual activations more

interpretable (Bricken et al., 2023), automatically searched for interpretations of them (Bills et al., 2023). Despite this progress, interpretability is a young field, and there are not yet many examples in which interpretability research has yielded findings that could meaningfully enhance an affirmative argument for safety. Future work could extend these approaches to identify concepts like "bioweapons" or "fraud" in AI activations.

4.2.2 Theoretical evidence on model internals

Explanation: Regulators could require formal and verifiable arguments that show that developers understand system internals.

While empirical evidence is valuable, model internals are highly complex and may be difficult to make arguments about. Theoretical approaches offer a way to overcome this hurdle because formal arguments can be automatically verified. Therefore, even as AI systems become more advanced and formal arguments are too complicated for developers to understand, such arguments can still be verified. Leveraging formal arguments requires two steps: (1) finding a statement which, if true, would provide evidence for an AI system's safety and (2) generating a proof of that statement.

Example: Eliciting latent knowledge. If developers could reliably determine what AI systems 'believe,' it would be much easier to trust and control them. For instance, developers could simply determine whether an AI system 'believes' that it is a good idea for humans to deploy it. However, for sufficiently-powerful models, human overseers may not be able to trust an AI system when it reports its beliefs. In the eliciting latent knowledge report, researchers try to examine if there are strategies that could guarantee that models reveal their true beliefs (Christiano et al., 2021). However, there are currently no known ways to guarantee that powerful models report their beliefs accurately. Some researchers have attempted to develop a system for making formal statements about model 'beliefs" (Christiano et al., 2022). While this approach is potentially promising, this research is in its early stages, and is not yet ready to be applied.

Related work: Formal verification of model behavior is an active ML research topic. The most common subproblem is 'certified robustness' – the problem of proving that a model's output will not change if inputs are perturbed by some small amount (Li et al., 2023). So far, there are few examples of formal guarantees providing evidence for safety outside of simple settings.

4.3 Development: Safe by design systems

Explanation: Regulators can require developers to provide formal verifications that powerful AI systems will behave safely within provable capability bounds.

There has been great interest in "safe by design" AI systems: systems that have formal guarantees based on mathematical or logical proofs. Safety by design can be applied at multiple steps throughout the development cycle. For example, proofs could be applied to model architectures (e.g., to show that a certain training process has guaranteeable safety properties), hardware (e.g., to show that hardware provably meets certain security requirements), code (e.g., to show that code meets certain criteria that suggests that it can be run safely even if it is not fully understood), and various other steps (see Tegmark & Omohundro, 2023).

Although some safe-by-design model architectures exist, current implementations of these ar-

chitectures are not performance-competitive with deep learning. Unfortunately, deep learning does not admit sufficiently tight safety bounds. In other words, deep learning algorithms are currently best at building powerful models, but they do not provide the kinds of formal safety arguments that we may achieve with "safe by design" architectures. This suggests that scientists may need to develop new "safe by design" architectures for sufficiently-powerful models, or they may need to sufficiently improve our understanding of the science of deep learning.

Example: Researchers develop a new paradigm with theoretical and mathematical guarantees. This paradigm is competitive with deep learning (i.e., it allows us to cost-effectively build powerful models) or it becomes clear that models past a certain capabilities threshold should only be designed using the safe-by-design architectures.

Related work: Some researchers are investigating safe-by-design architectures that could scale toward artificial general intelligence. Examples include approaches focused on mathematical proofs (e.g., Dalrymple, 2023), infra-bayesian physicalism (Kosoy, 2023), and proof-carrying code (Tegmark & Omohundro, 2023). Proof-carrying code could lead to automated software verification: mathematical proofs could be applied to code to guarantee that the code meets certain desired specifications. This approach could be necessary to verify that AI-generated code is safe to execute (see Tegmark & Omohundro, 2023).

5 Operational practices for affirmative safety

While our focus in this paper is on describing the technical components of affirmative safety, it is important to recognize that **operational practices** also play an important role. By "operational practices", we refer to aspects of an organization's culture, decision-making processes, and internal governance mechanisms that may increase or decrease certain kinds of risks. Three examples include **information security practices**, **safety culture**, and **emergency response capacity**.

Information security. Poor information security could lead to malicious actors stealing the weights of powerful AI systems. As a result, to show that an organization is keeping risks below acceptable risk thresholds, they may need to show that they have sufficient safeguards in place to protect their model weights (and other sensitive material that could allow malicious actors to create dangerous AI systems). This principle is already present in Anthropic's Responsible Scaling Policy: Anthropic publicly committed to not develop "ASL-3 systems" (AI that could substantially increase the risk of catastrophic misuse, for example by enabling large-scale biological attacks) until its information security standards were sufficiently strong "such that non-state attackers are unlikely to be able to steal model weights and advanced threat actors (e.g., states) cannot steal them without significant expense" (Anthropic, 2023). Ideally, information security standards would be checked by appropriate government red-teamers and enforced across the board.

Safety culture. Safety culture is commonly assessed in the realm of nuclear security. The International Atomic Energy Agency (IAEA) conducts safety culture assessments to review the culture of nuclear facilities and identify potential improvements (IAEA, 2016). Operational Safety Review Teams (OSART), consisting of international experts with experience in nuclear safety, conduct these assessments. The assessments include on-site evaluations (observations of operating procedures, review of relevant documents), interviews and surveys with staff, and an examination of the organization's decision-making track record. This process is used to assess

several aspects of safety culture; examples include leadership's commitment to safety, safety training, communication processes, risk management procedures, attitudes toward safety, risk reporting systems, employee understanding of risks, and allocation of resources for safety. An affirmative case for safety could require organizations to provide evidence of their safety culture or receive sufficiently high scores on safety culture assessments conducted by independent parties.

Emergency response capacity. Risks from advanced AI may arise suddenly and with short notice. As a result, an affirmative safety case may require institutions developing advanced AI to show that they have sufficient measures in place to detect and manage sudden risks. This principle is present in OpenAI's preparedness framework (OpenAI, 2023c): OpenAI safety researchers can "fast-track" information to leadership if "a severe risk rapidly develops" (OpenAI, 2023c). To expand on this, governments could require advanced AI companies to have emergency response plans that notify not only senior leadership at the AI company but also relevant national security figures or AI experts in the US government. In the event of an imminent AI-related emergency, it would be essential for government officials to be notified and have the ability to intervene. Emergency response plans could also include "kill switches" that allow governments to swiftly halt a dangerous AI experiment or have a company withdraw access to a dangerous AI model (Miotti & Wasil, 2023; Wasil, 2023).

6 Comparison to NIST AI Risk Management Framework

The National Institute of Standards and Technology (NIST) released the Artificial Intelligence Risk Management Framework (AI RMF). The framework describes desired criteria for AI systems: they ought to be (a) **valid and reliable**, (b) **safe**, (c) **fair and unbiased**, (d) **secure and resilient**, (e) **transparent and accountable**, (f) **explainable and interpretable**, and (g) **privacy-enhanced** (NIST, 2022). NIST's work is intended to offer a framework that can help companies reason about risks and make voluntary commitments.

NIST's work differs from our recommendations in a few important ways. First, NIST's Risk Management Framework is entirely voluntary – companies are free to ignore its recommendations. NIST describes the Risk Management Framework as "regulation-agnostic" and notes that the framework is not meant to supersede regulations and laws (NIST, 2023a). Second, and relatedly, NIST does not assign risk tolerance– it does not specify the level of risk that is considered acceptable in various domains. NIST's work has valuably helped introduce a common language when discussing risk management, define desired criteria, and pave the way for voluntary commitments. However, NIST recognizes the limitations of voluntary approaches, and the NIST framework should not be a substitute for binding regulations.

Notably, the NIST AI Risk Management Framework does recognize that certain kinds of AI development could pose unacceptably high-risk levels. "In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed" (NIST, 2023a). We agree strongly with this principle. Ideally, this principle would be instantiated by a regulatory body that reviews technical evidence (such as the evidence described above) and non-technical evidence (such as an organization's safety culture and information security practices) to determine if risks can be sufficiently managed.

References

AI Safety Institute. (2023). *Introducing the AI Safety Institute*. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

Anthropic. (2023, September). *Anthropic's Responsible Scaling Policy*. https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., ... & Saunders, W. (2023). *Language models can explain neurons in language models.* https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., ... & Olah, C. (2023). *Towards monosemanticity: Decomposing language models with dictionary learning.* Transformer Circuits Thread, 2. https://transformer-circuits.pub/2023/monosemantic-features

Center for AI Safety. (2023). *Statement on AI Risk.* https://safe-ai.webflow.io/statement-on-ai-risk

China Law Translate. (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services* https://www.chinalawtranslate.com/en/generative-ai-interim/

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences* https://doi.org/10.48550/arXiv.1706.03741

Christiano, P., Cotra, A., & Xu, M. (2021). *Eliciting latent knowledge: How to tell if your eyes deceive you.* https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/

Christiano, P., Neyman, E., & Xu, M. (2022). *Formalizing the presumption of independence.* https://doi.org/10.48550/arXiv.2211.06738

Clymer, J., Baker, G., Subramani, R., & Wang, S. (2023). *Generalization Analogies: A Testbed for Generalizing AI Oversight to Hard-To-Measure Domains* https://doi.org/10.48550/arXiv.2311.07723

Dalrymple, D. (2023). *Mathematics and modelling are the keys we need to safely unlock transformative AI* https://www.aria.org.uk/wp-content/uploads/2023/10/ARIA-Mathematics-and-modelling-are-the-keys-we-need-to-safely-unlock-transformative-AI-v01.pdf

Feynman, R. P. (1986). Volume 2: Appendix F - *Personal Observations on Reliability of Shuttle.* Report of the Presidential Commission on the Space Shuttle Challenger Accident. NASA. https://www.nasa.gov/history/rogersrep/v2appf.htm

Gao, L., Schulman, J. & Hilton, J. (2022). *Scaling Laws for Reward Model Overoptimization* https://doi.org/10.48550/arXiv.2210.10760

International Atomic Energy Agency (IAEA). (2016). *Performing Safety Culture Self-Assessments.* Safety Reports Series No. 83. Vienna. https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1682_web.pdf

International Electrotechnical Commission. (2010). IEC 61508-1:2010 https://webstore.iec.ch/publication/5515

Kosoy, V. (2023). *The Learning-Theoretic Agenda: Status 2023* https://www.alignmentforum.org/posts/ZwshvqiqCvXPsZEct/the-learning-theoretic-agenda-status-2023

Li, L., Xie, T., & Li, B. (2023, May). *Sok: Certified robustness for deep neural networks.* In 2023 IEEE symposium on security and privacy (SP) (pp. 1289-1310). IEEE. https://doi.org/10.48550/arXiv.2009.04131

Loizos, C. (2023). *StrictlyVC in conversation with Sam Altman, part two (OpenAI)* [Video] https://youtu.be/ebjkD1Om4uw?si=vqqJNIw0ue81ruaa&t=1340

Miotti, A., & Wasil, A. (2023). *Taking control: Policies to address extinction risks from advanced AI.* https://doi.org/10.48550/arXiv.2310.20563

MIT Technology Review. (2023). *Video: Geoffrey Hinton talks about the "existential threat" of AI.* https://www.technologyreview.com/2023/05/03/1072589/video-geoffrey-hinton-google-ai-risk-ethics/

National Institute of Standards and Technology (NIST). (2022). *AI Risk Management Framework: Second Draft* https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

National Institute of Standards and Technology (NIST). (2023a, January). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

National Institute of Standards and Technology (NIST). (2023b, October 26). *U.S. Artificial Intelligence Safety Institute* https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

Nuclear Regulatory Commission. (1983). *Safety Goals for Nuclear Power Plant Operation* https://www.nrc.gov/docs/ML0717/ML071770230.pdf

OpenAI. (2023b, December 14). *Weak-to-strong generalization* https://openai.com/research/weak-to-strong-generalization

OpenAI. (2023c, December 18). *Preparedness Framework (Beta)* https://cdn.openai.com/openai-preparedness-framework-beta.pdf

Oversight of A.I.: Principles for Regulation: Hearing before the Judiciary Committee Subcommittee on Privacy, Technology, and the Law, U.S. Senate, 118th Congr. (2023). (testimony of Dario Amodei). https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf

Park, P.S., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. (2023). *AI Deception: A Survey of Examples, Risks, and Potential Solutions* https://doi.org/10.48550/arXiv.2308.14752

Prime Minister's Office. (2023, November 1). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023* https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-

2-november-2023

Tegmark, M., & Omohundro, S. (2023). *Provably safe systems: the only path to controllable AGI*. https://doi.org/10.48550/arXiv.2309.01933

The Logan Bartlett Show. (2023). *Anthropic CEO on Leaving OpenAI and Predictions for Future of AI* [Video]. YouTube. https://www.youtube.com/watch?v=gAaCqj6j5sQ&t=5885

Wang, K., Variengien, A., Conmy, A., Shlegeris, B. & Steinhardt, J. (2022). *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small* https://doi.org/10.48550/arXiv.2211.00593

Wasil, A. (2023) *Addressing Global Security Risks From Advanced AI* https://medium.com/fidutam/addressing-global-security-risks-from-advanced-ai-e81cc54d0c90

World Nuclear Association. (2022). *Safety of Nuclear Power Reactors* https://www.world-nuclear.org/information-library/safety-and-security/safety-of-plants/safety-of-nuclear-power-reactors.aspx

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., ... & Hendrycks, D. (2023). *Representation engineering: A top-down approach to AI transparency.* https://doi.org/10.48550/arXiv.2310.01405