ANTHROP\C

Written Statement for AI Insight Forum: Risk, Alignment, & Guarding Against Doomsday Scenarios

Dr. Jared Kaplan, Co-Founder and Chief Science Officer
Anthropic PBC

December 6, 2023

Introduction

Thank you for the opportunity to participate in the Senate AI Insight Forum on Risk, Alignment, & Guarding Against Doomsday Scenarios. I am Dr. Jared Kaplan, Anthropic's Co-Founder and Chief Science Officer.

Anthropic was founded in 2021 as an AI safety and research company to build reliable, interpretable, and steerable AI systems. We founded Anthropic because we believe the impact of AI might be comparable to that of the industrial and scientific revolutions. We also believe this level of impact could start to arrive soon—perhaps in the coming decade—and the benefits of AI will be truly profound. In the next few years, AI could greatly accelerate treatments for diseases such as cancer, lower the cost of energy, revolutionize education, improve efficiency throughout government, and much more. While we do not believe that the systems available today pose an imminent concern, we believe that we need to lay the groundwork now to ensure future, more powerful systems are safe.

I'd note that almost everyone who has said "the thing we're working on might be one of the biggest developments in history" has been wrong. However, based on the technological trends we see, we are concerned about how the rapid deployment of increasingly powerful AI systems will impact society in the short, medium, and long term, and we believe there is enough evidence to seriously prepare for a world where rapid AI progress leads to transformative AI systems. Our submission shares, at a high-level, Anthropic's approach to AI risk management and AI safety research. At Anthropic, our motto has been "show, don't tell," and we've focused on releasing a steady stream of safety-oriented research that we believe has broad value for the AI community. This research also informs our work on developing responsible AI policies and governance, which includes our recently released Responsible Scaling Policy. By conducting rigorous research on AI's implications today, we aim to provide policymakers and researchers with the insights and tools they need to help mitigate these potentially significant societal harms and ensure the benefits of AI are broadly and evenly distributed across society.

AI Safety Risks

The three main ingredients leading to predictable[1] improvements in AI performance are training data, computation, and improved algorithms. In the mid-2010s, we discovered that larger AI systems were consistently smarter, and so we theorized that the most important ingredient in AI performance might be the total budget for AI training computation. When this was graphed, it became clear that the amount of computation going into the largest models was growing at 10x per year. In 2019, several members of Anthropic's founding team made this idea precise by developing scaling laws for AI, demonstrating that you could make AIs smarter in a predictable way, just by making them larger and training them on more data. AI systems are now

[1] Algorithmic progress – the invention of new methods for training AI systems – is more difficult to measure, but progress appears to be exponential and faster than Moore's Law. When extrapolating progress in AI capabilities, the exponential growth in spending, hardware performance, and algorithmic progress must be multiplied in order to estimate the overall growth rate.

approaching human-level performance on a large variety of tasks, and yet training these systems still costs far less than “big science” projects like the Hubble Space Telescope or the Large Hadron Collider—meaning that there’s a lot more room for growth.

With this in mind, there are two commonsense reasons to be concerned. First, it may be tricky to build safe, reliable, and steerable systems when those systems are starting to become as intelligent and as aware of their surroundings as their designers. Second, rapid AI progress would be very disruptive, changing employment, macroeconomics, and power structures both within and between nations. That is why we believe it is necessary to err on the side of caution.

The Role of Frontier Models in Empirical Safety

A major reason Anthropic exists as an organization is because we believe that safety research on “frontier” AI systems is necessary. Conducting this safety research requires an institution that can both work with large models and prioritize safety.[2] Since many of our most serious safety concerns might only arise with near-human-level systems, we need to understand how safety methods and properties change as models scale. If future large models turn out to be very dangerous, it’s essential we develop compelling evidence this is the case. We expect this to only be possible by using large models.

Unfortunately, if empirical safety research requires large models, that forces us to confront a difficult trade-off. We must make every effort to avoid a scenario in which safety-motivated research accelerates the deployment of dangerous technologies. But we also cannot let excessive caution make it so that the most safety-conscious research efforts only ever engage with systems that are far behind the frontier, thereby dramatically slowing down vital research. Furthermore, we think that in practice, doing safety research isn’t enough—it’s also important to build an organization with the institutional knowledge to integrate safety research into real systems as quickly as possible. Navigating these tradeoffs responsibly is a balancing act, and these concerns are central to how we make strategic decisions as an organization.

Anthropic’s Current Safety Research

At Anthropic, our AI safety research is grounded in empirical evidence. We are working in a variety of directions to discover how to train safe AI systems. Some key ideas currently under development include:

- **Mechanistic Interpretability**. Mechanistic interpretability is the project of trying to reverse engineer neural networks into human understandable algorithms, similar to how one might reverse engineer an unknown and potentially unsafe computer program. Our hope is that this may eventually enable us to do something analogous to a “code review,” auditing our models to either identify unsafe aspects or else provide strong guarantees of safety. We believe this is a very difficult problem, but we are optimistic that it is tractable. We

[2] Effective safety research on large models doesn't just require nominal access to these systems – to do work on interpretability, fine tuning, and reinforcement learning it’s necessary to develop AI systems internally at Anthropic.

understand significantly more about the mechanisms of neural network computation than we did even a year ago, such as those responsible for memorization and copying.

- **Scalable Oversight**. The main goal of scalable oversight is to get models to better understand and behave in accordance with human values. Turning language models into aligned AI systems will require significant amounts of high-quality feedback to steer their behaviors. However, a major concern is that humans, for a number of reasons, may not be able to provide models with the necessary high-quality feedback.[3] That is why we believe the only way to provide models with the necessary supervision will be to have AI systems partially supervise themselves or assist humans in their own supervision. To do this effectively, we need to magnify a small amount of high-quality human supervision into a large amount of high-quality AI supervision. This idea is already showing promise through techniques such as reinforcement learning from human preferences (RLHF)[4] and Constitutional AI,[5] though we see room for much more to make these techniques reliable with human-level systems. Another key feature of scalable oversight techniques is that they allow us to automate red-teaming—meaning that we can automatically generate potentially problematic inputs to AI systems, see how they respond, and then automatically train them to behave in ways that are more honest and harmless. The hope is that we can use scalable oversight to train more robustly safe systems.

- **Learning Processes Rather than Achieving Outcomes**. Many of the concerns about the safety of advanced AI systems are addressed by training these systems in a process-oriented manner. In process-oriented learning, the goal is not to achieve the final outcome but to master individual processes that can then be used to achieve that outcome. Limiting AI training to process-oriented learning might be the simplest way to improve a host of issues with advanced AI systems. We are also excited to identify and address the limitations of process-oriented learning, and to understand when safety problems arise if we train with mixtures of process and outcome-based learning. We currently believe process-oriented learning may be the most promising path to training safe and transparent systems up to and somewhat beyond human-level capabilities.

- **Understanding Generalization**. LLMs have demonstrated a variety of surprising emergent behaviors, from creativity to self-preservation to deception. While all of these behaviors likely arise from the training data, the pathway is complicated: the models are first “pretrained” on gigantic quantities of raw text, from which they learn wide-ranging representations and the ability to simulate diverse agents. Then they are fine-tuned in myriad ways. We are working on techniques to trace a model’s outputs back to the training data to yield an important set of cues for understanding model behavior.

---

[3] Irving & Askell, “AI Safety Needs Social Scientists,” (Feb. 19, 2019), *available at*: https://distill.pub/2019/safety-needs-social-scientists/ (Accessed Nov. 28, 2023).
[4] Christiano, Leike, Brown, Martic, Legg & Amodei, “Deep Reinforcement Learning from Human Preferences,” *available at*: https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf (Accessed Nov. 28, 2023).
[5] Anthropic, Constitutional AI: Harmlessness from AI Feedback, (Dec. 15, 2022), *available at*: https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback (Accessed Nov. 28, 2023).

- **Testing for Dangerous Failure Modes**. One key concern is the possibility that an advanced AI may develop harmful emergent behaviors, such as deception or strategic planning abilities. We think the way to anticipate this kind of problem before it becomes a direct threat is to set up environments where we deliberately train these properties into small-scale models that are not capable enough to be dangerous, so that we can isolate and study them. We aim to build detailed quantitative models of how these tendencies vary with scale so that we can anticipate the sudden emergence of dangerous failure modes.

- **Societal Impacts and Evaluations**. Critically evaluating the potential societal impacts of our work is a key pillar of our research. Our approach centers on building measurements to evaluate and understand the capabilities, limitations, and potential for the societal impact of our AI systems. For example, we have published research analyzing predictability and surprise in large language models, which studies how the high-level predictability and unpredictability of these models can lead to harmful behaviors. In that work, we highlight how surprising capabilities might be used in problematic ways. We have also studied methods for red-teaming language models to discover and reduce harms by probing models for offensive outputs across different model sizes.

Anthropic's Responsible Scaling Policy

In September, Anthropic published our Responsible Scaling Policy (RSP)[6], a series of technical and organizational protocols that we are adopting to help us manage the risks of developing increasingly capable AI systems. Our RSP focuses on managing catastrophic risks—those where an AI model has the potential to directly cause large-scale devastation. Such risks can come from deliberate misuse (e.g., use by terrorists or state actors to create bioweapons) or from models that cause destruction by acting autonomously in ways contrary to the intent of their designers.

Our RSP defines a framework called AI Safety Levels (ASL) for addressing catastrophic risks, modeled loosely after the U.S. government's biosafety level (BSL) standards for handling of dangerous biological materials. The basic idea is to require safety, security, and operational standards appropriate to a model's potential for catastrophic risk, with higher ASL levels requiring increasingly strict demonstrations of safety. Our intention is for the ASL system to strike a balance between effectively targeting catastrophic risk and incentivising beneficial applications and safety progress. On the one hand, the ASL system implicitly requires us to temporarily pause training of more powerful models if our AI scaling outstrips our ability to comply with the necessary safety procedures. But it does so in a way that directly incentivizes us to solve the necessary safety issues as a way to unlock further scaling, and allows us to use the most powerful models from the previous ASL level as a tool for developing safety features for the next level.[7] If adopted as a

[6] https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf
[7] As a general matter, Anthropic has consistently found that working with frontier AI models is an essential ingredient in developing new methods to mitigate the risk of AI.

standard across frontier labs, we hope this might create a “race to the top” dynamic where competitive incentives are directly channeled into solving safety problems.

We have been encouraged by feedback from senior officials in the United Kingdom calling for other AI developers to propose RSP-like mechanisms of their own.[8] We believe RSPs can serve as a valuable prototype and test case for potential downstream regulation. We expect to learn much more about what sensible risk-assessment and mitigation could look like by empirically testing ideas, and are excited to share what we learn with policymakers. We hope that by adopting early iterations of these policies, we will generate evidence about ways to turn safety from a high-level concept into something implemented in technical and process-driven ways in fast-moving technical organizations.

Conclusion

We believe that artificial intelligence may have an unprecedented impact on the world, potentially within the next decade. We want to be clear that we do not believe that the systems available today pose an imminent concern. However, it is prudent to do foundational work now to help reduce risks from advanced AI if and when much more powerful systems are developed. It may turn out that creating safe AI systems is easy, but we believe it’s crucial to prepare for less optimistic scenarios. Anthropic remains committed to taking an empirically driven approach to AI safety and to implementing risk management measures through our RSP that can help us succeed across a range of scenarios.

[8] “U.K. aiming for global agreement on dangerous AI,” Vincent Manancourt. *POLITICO Pro*, Sept. 25, 2023. (Accessed Nov. 28, 2023)