Malo Bourgon
Chief Executive Officer
Machine Intelligence Research Institute
Berkeley CA

November 29, 2023

U.S. Senate AI Insight Forum:
Risk, Alignment, & Guarding Against Doomsday Scenarios

Written statement

Leader Schumer, Senator Rounds, Senator Heinrich, and Senator Young, thank you for the invitation to participate in the AI Insight Forum series, and for giving me the opportunity to share the perspective of the Machine Intelligence Research Institute (MIRI) on the challenges humanity faces in safely navigating the transition to a world with smarter-than-human artificial intelligence (AI).

MIRI is a research nonprofit based in Berkeley, California, founded in 2000. Our focus is forward-looking: we study the technical challenges involved in making smarter-than-human AI systems safe.

To summarize the key points I'll be discussing below: (1) It is likely that developers will soon be able to build AI systems that surpass human performance at most cognitive tasks. (2) If we develop smarter-than-human AI with anything like our current technical understanding, a loss-of-control scenario will result. (3) There are steps the U.S. can take today to sharply mitigate these risks.

It is likely that developers will soon be able to build AI systems that surpass human performance at most cognitive tasks

The goal of the field of AI, going back to its inception in 1956, is to reproduce in machines the reasoning and problem-solving abilities humans possess—our "intelligence." The reason for pursuing this line of research is obvious: human-style intelligence is extremely powerful. The largest historical improvements in human welfare have generally been the product of humans applying our intelligence to solve novel problems and develop new technologies. Creating truly intelligent machines, if we can align them with our interests, would allow us to effectively automate this process of discovery and technological development, allowing us to create a radically more prosperous future.

Until recently, AI systems were very *narrow*, in the sense that a given AI system could only perform a specific task or type of cognitive labor, like playing chess. However, in the last few years, advances in both software and hardware have enabled generative AI systems, including large language models (LLMs), to become much more *general* in their intelligence. State of the art LLMs, such as ChatGPT, can carry on a conversation, write poetry, ace many standardized tests, translate between languages, write computer programs, etc. These systems are far from perfect, and often make silly mistakes, but they are enormously more capable than the best AI systems from a year or two prior, and are continuing to improve rapidly.

The impacts of current AI systems, both positive and negative, are already visible in a lot of Americans' daily lives, my own included. My wife runs her own digital graphic design asset business, and it's looking more and more like her business may crash soon with the advent and rapid improvement of AI image generators. You've no doubt heard similar stories from many others, and as AI systems continue to become increasingly capable, their disruptive effects on society will only continue to increase.

Forecasting technical progress is notoriously difficult. It's challenging to predict how many years away we are from AI systems that can match humans at most cognitive tasks. A growing number of AI researchers in academia and at the leading AI labs think it's increasingly likely that general human-level AI systems will be created within the next decade, and potentially within a few years.

Importantly, we should not expect AI progress to stop at roughly human-level abilities. If we succeed at creating general artificial intelligence at all, we should also expect to see AI systems progress far beyond human-level.

When scientists automate a cognitive task, e.g., playing chess, what we've generally found is that AI quickly and permanently surpasses human-level abilities at that task. Humans are nowhere near a cognitive ceiling, and while we're "smart" compared to other species, AI keeps crushing us at the tasks we automate, once we figure out how to automate them at all. Thinking ahead, we should expect the same to happen when we automate critical tasks like "advancing fields of science," "building out new technologies," "persuading and manipulating others," and "strategizing about how to gain power in the larger world."

All of these are just more tasks we can automate, once the technology reaches that point; but these are tasks of monumental consequence.

If we develop smarter-than-human AI with anything like our current technical understanding, a loss-of-control scenario will result

Geoffrey Hinton, one of the three godfathers of the current AI paradigm, deep learning, has previously noted that "there is not a good track record of less intelligent things controlling things of greater intelligence." This is a succinct distillation of the source of the risk we may soon face. Indeed, our dominance as a species, driven by our relatively superior intelligence, has led to a mass extinction event of other species over the last 10,000 years—not because we bear any hostility towards them, but because we have a variety of goals and these goals happen to conflict with their interests.

To successfully navigate the transition to a world with smarter-than-human AI, we need the ability to direct, or *align*, AI systems to want what we want, to support a positive future. To do that, we'll need, at a minimum, a deep understanding of how these systems work.

Unfortunately, we currently have no real understanding of how they work, or why they do what they do. This is a result of the fact that current machine learning looks less like building AIs, and more like "growing" them. Quoting Hinton again, from a 60 Minutes interview:

> Interviewer: What do you mean, "we don't know exactly how [AI] works"? It was designed by people.
>
> Hinton: No, it wasn't. What we did was we designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things, but we don't really understand exactly how they do those things.

Advanced deep learning-based AI systems are akin to black boxes. We can see what we put in, and what comes out, but when we look inside the box all we see are trillions of fractional numbers, that somehow produce the output we observe from the system.

AI research and deployment are progressing at a dizzying pace—yet after over a decade of research attempting to figure out how to align such systems, it's MIRI's assessment that alignment is a surprisingly difficult technical problem that humanity is nowhere near solving. We remain hopeful that alignment is solvable in principle, and would eventually be solved given a large (plausibly multi-generational) research effort; but time seems to be running short, and the most likely outcome of developing smarter-than-human AI prior to solving alignment is human extinction.

Present-day AI systems do not pose an existential threat, but there is a significant chance that systems in the near future will, as they become capable of performing increasingly complex multi-step tasks without humans in the loop.

There are steps the U.S. can take today to sharply mitigate these risks

Soon after it becomes possible for AI developers to create smarter-than-human AI, the likely outcome is loss-of-control, unless there are robust global checks on AI development. These checks should ensure that smarter-than-human AI systems are not developed until there is a broad scientific consensus that we fully understand the relevant phenomena and it is safe to proceed; and until we are able to secure these systems such that they can't fall into the hands of malicious or incautious actors.[1]

This is a tall order that will likely require unprecedented feats of global coordination, that will entail confronting and navigating many challenging tradeoffs. As the global leader in AI, the U.S. has the unique ability and responsibility to lead by example and establish *domestic AI regulation*, while collaborating with world governments to create a *global AI coalition*. These policies can be enforced by *governing computing hardware*.

Domestic AI regulation: The U.S. government needs the capacity to rapidly identify and respond to AI threats. Ideally, we would create a new office with the goal of monitoring and responding to AI risks, with the power to license cutting edge or frontier AI development, commensurate with the risks as the technology evolves over time. This will require a significant increase in the technical expertise in AI risk analysis within government, and interagency collaboration.

In the meantime, we could require that companies perform substantial evaluation for dangerous capabilities and report the results to the recently announced AI Safety Institute (AISI) within the Department of Commerce.[2] If the AISI identifies a national security emergency as a result of these reported capabilities, it could be empowered to stop the development of AI systems which pose an unacceptably high risk to national security and the global community. Congress could also clarify when companies at different stages of AI development and application are liable for damages caused by a system, and could establish that the duty of care for AI developers includes rigorous and accurate evaluations.

Separately, in order to build out AI expertise in government and provide a test bed for AI system risk analysis, the U.S. government could fund additional research into model evaluations for dangerous capabilities, either by increasing funding for NIST, specifically allocating part of the NAIRR, or creating a public-private third party auditing regime. This research could lay the groundwork for both unearthing and regulating future threats.

Some AI companies have voluntarily adopted risk mitigating policies, such as Anthropic with their Responsible Scaling Policy. Unfortunately, the explosion in valuation of "AI companies" shows a nigh-insurmountable financial incentive to continue developing more capable models even when AI systems pose severe risks. Even without such massive incentives, existing voluntary commitment processes are too nascent to prevent catastrophic risks or loss-of-control within the short-term. The government should therefore step in to require sufficient safeguards.

[1] AI company executives have acknowledged that they are currently unable to secure AI systems from highly motivated state actors, a view backed up by independent analysis.

[2] The recent AI Executive Order requires that AI developers report the results of evaluations that they run, but the wording is ambiguous about whether developers are required to run evaluations in the first place. Merely requiring reporting would obviously be insufficient, as this would provide a disincentive to run evaluations in the first place.

Global AI Coalition: Preventing the development of smarter-than-human AI systems until adequate safety techniques are developed likely requires that the U.S. spearhead an international effort to secure AI computing hardware and frontier AI development efforts. International coordination is critical to ensure that allied countries can regulate or pause AI without just driving AI development to foreign competitors.

Due to America's clear leadership in AI capabilities, we can and should lead the world towards an international coalition for safe AI. Countries that join this coalition would agree to not build, or allow to be built, smarter-than-human AI until a scientific consensus forms that it is safe to do so. The U.S. and allies should expand the use of hardware export controls, to prevent countries outside of this coalition, from importing or building production capacity for advanced chips, making it essentially impossible to train such systems.

Congress can lay the foundation for adequate international oversight of frontier AI development now by:

1. Studying the impact and efficacy of requiring or incentivizing powerful ML chips to be kept in large cloud data centers located in places where it is possible to enforce rules on their usage.
2. Funding research into compliance techniques that could verify whether a given compute cluster is being used for AI development.

Governing Computing Hardware: Compute provides a key enforcement lever because computational power is the main difficult-to-obtain ingredient that has driven the explosive growth in general AI capabilities development. It is currently very difficult to build state-of-the-art AI systems without expensive hardware, and it is very difficult to prevent actors who have sufficient access to this hardware from developing powerful AI systems.

Congress can promote the security of compute hardware used to train AI systems by:

1. Establishing a registry of advanced AI chips that tracks basic information like location, ownership, use case, and cluster size.
2. Increasing the budget of the Department of Commerce to better enforce existing hardware export controls to reduce chip smuggling and other loopholes.
3. Directing and supporting the executive to explore an international alliance to restrict all frontier AI hardware to a limited number of large computer clusters, and placing those clusters under a uniform monitoring regime to forbid uses that endanger humanity; offering symmetrical treatment to signatory countries, and not permitting exceptions to any governments or militaries; in order to retain the option of refusing dangerous AI pathways as a united people.