Opening statement (as delivered) before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, hearing on "Oversight of A.I.: Principles for Regulation"

July 25th, 2023

Stuart Russell
Professor of Computer Science
The University of California, Berkeley

Thank you, Chair Blumenthal and Ranking Member Hawley, and members of the Subcommittee, for the invitation to speak today, and for your excellent work on this vital issue.

AI, as we all know, is the study of how to make machines intelligent. Its stated goal is *general-purpose artificial intelligence*, sometimes called AGI or artificial general intelligence: machines that match or exceed human capabilities in every relevant dimension.

The last 80 years have seen a lot of progress towards that goal. For most of that time, we created systems whose internal operations we understood, drawing on centuries of work in mathematics, statistics, philosophy, and operations research.

Over the last decade, that has changed. Beginning with vision and speech recognition, and now with language, the dominant approach has been end-to-end training of circuits with billions or trillions of adjustable parameters. The success of these systems is undeniable, but their internal principles of operation remain a mystery. This is particularly true for the large language models or LLMs, such as ChatGPT.

Many AI researchers now see AGI on the horizon. In my view, LLMs do not *constitute* AGI, but they are a piece of the puzzle. We're not sure what shape the piece is yet or how it fits into the puzzle, but the field is working hard on those questions, and progress is rapid.

If we succeed, the upside could be enormous: I've estimated a cash value of at least 14 quadrillion dollars for this technology—a huge magnet in the future, pulling us forward.

On the other hand, Alan Turing, the founder of computer science, warned in 1951 that once AI "outstrips our feeble powers … we should have to expect the machines to take control."

We have pretty much completely ignored this warning. It's as if an alien civilization warned us by email of its impending arrival, and we replied, "Humanity is currently out of the office." Fortunately, humanity is now back in the office and has read the email from the aliens.

Of course, many of the risks from AI *are* well recognized already, including bias, disinformation, manipulation, and impacts on employment. I'm happy to discuss any of these.

But most of my work over the last decade has been on the problem of *control*: how do we maintain power, *forever*, over entities more powerful than ourselves?

The core problem we have studied comes from AI systems pursuing fixed objectives that are misspecified—the so-called King Midas problem. For example, social media algorithms were trained to maximize clicks and learned to do so by manipulating human users and polarizing societies. But with LLMs we don't even know what their objectives are. They learn to imitate humans and probably absorb all-too-human goals in the process.

Now, regulation is often said to stifle innovation, but *there is no real tradeoff between safety and innovation*. An AI system that harms human beings is simply not good AI. And I believe analytic predictability is as essential for safe AI as it is for the autopilot on an airplane.

This committee has discussed ideas such as third-party testing, licensing, a national agency, and an international coordinating body, all of which I support. Here are some more ways to, as it's said, "move fast and fix things":

- First, an absolute right to know if one is interacting with a person or a machine.
- Second, no algorithms that can decide to kill human beings, particularly when attached to nuclear weapons.
- Third, a kill switch that must be activated if systems break into other computers or replicate themselves.
- Fourth, go beyond the voluntary steps announced last Friday: systems that break the rules must be recalled from the market, for anything from defaming real individuals to helping terrorists build biological weapons.

Now developers may argue that preventing these behaviors is too hard, because LLMs have no notion of truth and are just trying to help. This is no excuse.

Eventually—and the sooner the better, I would say—we will develop forms of AI that are *provably safe and beneficial*, which can then be mandated. Until then, we need real regulation and a pervasive culture of safety.

Thank you.