Written Statement of Stuart Russell, Professor of Computer Science, UC Berkeley
For the United States Senate AI Forum on
"Risk, Alignment, & Guarding Against Doomsday Scenarios"
December 6th, 2023

Thank you for the opportunity to contribute to this forum; and thank you to the Senate for its serious and bipartisan engagement with the important questions arising from recent advances in AI.

The following remarks draw in part on my written testimony to the U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, & the Law, on July 25th, 2023.

I am primarily an AI researcher, with over 40 years of experience in the field. I am motivated by the potential for AI to amplify the benefits of civilization for all of humanity. My research over the last decade has focused on the *problem of control*: how do we maintain power, forever, over entities that will eventually become more powerful than us? How do we ensure that AI systems are safe and beneficial for humans? These are not purely technological questions. In both the short term and the long term, regulation has a huge role to play in answering them.

Executive summary

- The long-standing goal of AI has been to create *general-purpose AI systems*—also known as artificial general intelligence or AGI—that match or exceed human capabilities in every relevant dimension.
- Progress on AI capabilities over the last decade has been extremely rapid and is accelerating. Many researchers feel that AGI is on the horizon; some say it already here in rudimentary form.
- Handled well, the economic value of AGI would be enormous. This is already creating massive investment flows, which will increase as the goal gets closer.
- Given our current lack of understanding of how to control AGI systems and to ensure with absolute certainty that they remain safe and beneficial to humans, achieving AGI would present catastrophic risks to humanity, up to and including human extinction.
- It is essential to create a regulatory framework capable of adapting to these increasing risks while responding to present harms. The onus should be on developers to demonstrate rigorously to regulators that their systems are safe before they can be released.

Background: The Nature of Artificial Intelligence

The field of artificial intelligence aims to design machines capable of intelligent behavior. From its earliest days, the stated goal has been *general-purpose artificial intelligence*, sometimes called AGI or artificial general intelligence: machines that match or exceed human capabilities in every relevant dimension.[1]

To explicate the notion of "intelligent behavior," AI researchers have borrowed from definitions of *rationality* in philosophy and economics: roughly speaking, machines are intelligent to the extent that their actions can be expected to achieve their objectives. The objectives that machines pursue are

[1] The relevant dimensions *do not include sentience*, about which AI has little to say. Many films such as *Terminator*, *Ex Machina*, and *Mission Impossible: Dead Reckoning* identify the unexpected emergence of consciousness in machines is the primary concern—a misunderstanding repeated in thousands of media articles about AI. In fact, competence is the problem, just as it is for a human chess player losing to a more competent chess program.

typically provided by us: for example, we define checkmate in chess and design algorithms that pursue it; we tell the navigation app our destination and it finds a way to reach it. In other words, we build objective-achieving machines, we feed objectives into them or specialize them for particular objectives, and then the machines do the rest.

For most of its history, AI has been analytical in its approach: breaking down intelligence into its constituent parts, understanding and implementing each part in mathematical and computational terms, and combining the parts to create functioning intelligent systems. This process of deliberate, component-based, mathematically rigorous design made AI similar in many ways to other branches of engineering such as aeronautics, electronics, and nuclear engineering. By and large, the behavior of AI systems was predictable, and it was usually possible to predict in advance whether a given design modification would result in improved performance.

Over the last decade, with the advent of deep learning, that has changed. Beginning with vision and speech recognition, and now with language, the dominant approach has been end-to-end training of "deep neural networks"—essentially circuits with billions or trillions of adjustable parameters. The training consists of quintillions (or more) of small random adjustments to the parameters to improve the circuit's performance on vast data sets. These methods have led to roughly human-level performance in many important tasks, including speech recognition, machine translation, and object recognition in images. Large language models or LLMs, such as ChatGPT, are considered by some to be the first examples of general-purpose AI by virtue of the fact that human language itself is general-purpose and LLMs can be trained on almost the entirety of human knowledge as recorded in textual form.

Once trained, deep learning systems perform well, but their internal principles of operation remain a mystery. They are black boxes—not because we cannot examine their internals, but because their internals are largely impossible to understand. This is particularly true for LLMs, which may have hundreds of billions or even trillions of parameters. Furthermore, LLMs are trained not to achieve specific objectives but to imitate humans; one expects that this leads them to have human-like objectives, but we cannot determine what those are or when they might be activated. Many labs are experimenting with building agents—systems capable of planning and acting in the real world—using LLMs as components in various ways.

Risks from current AI systems

A number of risks from existing AI systems have been studied extensively, including biased and opaque decisions affecting people's rights; manipulation by social media recommender systems that gradually modify people's preferences and personalities in order to turn them into more predictable consumers of content; disinformation (including deepfakes) and its use in foreign influence campaigns, seriously threatening our democracy; and the impact of AI on jobs in many sectors, accelerating the "hollowing out" of the American economy. These risks are serious, and my Senate testimony from the July 25th hearing outlined potential solutions, several of them requiring legislative action.

New categories of risk are materializing on an almost weekly basis, as new capabilities come to the fore. A brief summary follows, with more details and citations given in my July 25th written testimony:

- *Biosecurity risk* arises from the ability of AI systems to generate or disseminate knowledge related to the synthesis of toxins and disease organisms. For example, a recent paper shows that an AI system designed for pharmaceutical drug discovery could be repurposed trivially to propose thousands of new, highly toxic compounds. Another study conducted with students at

MIT showed how easy it is to get detailed advice from LLMs on creating new pathogens, despite "guardrails" intended to prevent exactly this.

- *Hallucination risk* arises from LLMs generating plausible, authoritative outputs that are completely fabricated. Significant examples include fabricated medical literature and nonexistent legal precedents, as well as defamation of real individuals. LLMs are also capable of inducing a form of hallucination in their users: millions of people have been seduced into relying on LLMs as their primary emotional contact, leaving them vulnerable to software updates that undermine their imagined connection.
- *Cybersecurity risk* arises from LLMs' ability to write and execute computer programs, from their extensive knowledge of cyberattack methods, and from their social manipulation abilities. At the Bletchley Park AI Summit, a member of the UK AI Safety Institute showed an LLM conducting a phishing attack against a (fictional) Harvard student. The LLM wrote and then sent a letter indicating problems with registration and payment of fees; the letter contained a link to an official-looking Harvard login page designed to capture passwords.
- *Manipulation risk* with LLMs differs from the corresponding risk with social media algorithms. As noted above, LLMs may acquire human-like goals, including goals of persuasion, as well as a "point of view" on particular topics. If hundreds of millions of people are using chatbots on a daily basis, that could have a significant and unpredictable impact on public opinion in any area. For example, it might lead to a gradual increase in hostile attitudes towards China, making a nuclear war more and more likely for no good reason. The possibility that opposite persuasion goals—for example, for and against climate-related policies – can be activated by different people in their interactions also leads to a polarization risk.

These risks are all *dramatically increased* by the availability of open-source LLMs, for two obvious reasons. First, problems of misuse of existing LLMs are exacerbated by making them freely available to bad actors. For example, it has been estimated that the 2016 Russian disinformation campaign against the US cost roughly $20 million, but would now cost closer to $1000.[2] Second, despite extensive efforts to train LLMs not to answer certain types of questions, it is trivially easy to remove these "guardrails" from open-source models.[3]

Prospects for general-purpose AI

The quest for AGI is accelerating. Ian Hogarth, Chair of the UK Government's Foundation Model Taskforce, reports a 100-million-fold increase since 2012 in compute budgets for the largest machine learning projects and "eight organizations raising $20bn of investment cumulatively in [the first three months of] 2023" for the express purpose of developing AGI. This amount is approximately ten times larger than the entire budget of the US National Science Foundation for the same period.[4]

Some experts view AGI as imminent. For example, a distinguished team of researchers at Microsoft who spent several months evaluating GPT-4 claimed that it shows "sparks of artificial general intelligence."[5] In a recent interview, Google Deepmind CEO Demis Hassabis stated, "In maybe a couple of years' time

---

[2] Lyric Jain, testimony to the UK House of Lords Communications and Digital Committee, September 19, 2023.
[3] See, for example, Lermen et al., LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B, or Bailey et al., Image Hijacks: Adversarial Images can Control Generative Models at Runtime.
[4] Ian Hogarth, "We must slow down the race to God-like AI," *Financial Times*, April 13, 2023.
[5] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y., "Sparks of Artificial General Intelligence: Early experiments with GPT-4," arXiv:2303.12712, 2023.

today's chatbots will look trivial by comparison to I think what's coming in the next few years. … I would not be surprised if we approached something like AGI or AGI-like in the next decade."[6]

Such claims are hard to evaluate. We have no understanding of LLM's internal principles of operation, and our intuitions are untrustworthy because we lack experience with entities that have read and absorbed (in some sense) thousands of times more text than any human being. What may appear to be an entirely original answer may in fact result from blending and mapping existing answers from a range of "nearby" sources.

In my view, LLMs are probably a piece of the AGI puzzle, but we do not yet know the shape of the piece and what other pieces are needed to complete the puzzle. Complacency is not advisable, however, because, as noted above, many research groups are looking for ways to construct entire functioning agents using LLMs as components. Every single AI researcher I have spoken to in the last year now views AGI as much closer than previously estimated. It is far better to prepare now and then find we have plenty of time to spare, than to prepare too late and find our species at a dead end.

Potential risks of general-purpose AI

Without wishing to imply that the harms from existing AI systems should be neglected, I believe it is essential to give adequate attention to the problem of control: how do we maintain power, forever, over entities that will eventually become more powerful than us? Many AI researchers and industry leaders have signed the following statement:[7]

> "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

Alan Turing, the founder of computer science, viewed the problem as insoluble:[8]

> "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control."

Within the standard model of AI, the most obvious failure mode is the *King Midas problem*: AI systems pursuing fixed objectives that are misspecified. Social media recommender systems provide an early example of this: in trying to maximize the clickthrough or engagement objective, they learn to manipulate humans and polarize societies. These are very simple algorithms, of course. More intelligent AI systems can take steps to preempt human interference, acquire additional resources, and (if necessary) deceive humans about their intentions, all in the service of a given objective. The literature on AI safety contains many scenarios illustrating the process whereby humans lose control in this way.[9] As noted above, the situation with LLMs is worse: we don't even know what their objectives are. They are simply trained to imitate humans, and they may absorb all-too-human goals in the process.

---

[6] Nilay Patel, "Inside Google's big AI shuffle — and how it plans to stay competitive," *The Verge*, July 10, 2023.
[7] Center for AI Safety, "Statement on AI Risk," May 30, 2023.
[8] Alan Turing, "Intelligent machinery, a heretical theory," a lecture given to the 51 Society, Manchester, 1951. Typescript available at turingarchive.org.
[9] Stuart Russell, *Human Compatible*, Viking, 2019; Nick Bostrom, *Superintelligence*, Oxford University Press, 2014; Max Tegmark, *Life 3.0*, Knopf, 2017, and Andrew Critch and Stuart Russell, "TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI," arXiv:2306.06924, 2023.

It is important to note that an AI system need not have physical embodiment and built-in weapons to have an enormous negative impact. AI systems are already empowered to send email, post on social media, purchase goods and services online (including real-world physical services such as DNA synthesis), and hire humans to carry out any task. The emergence of fully automated online corporations (e.g., trading or lending operations, language- or image-based services) is expected soon, and these will gradually extend their operations into the physical world through proxies. AI systems may also be able to gain access to remotely piloted and fully autonomous weapons, missile systems, satellite data feeds, etc.

In summary, **further development towards AGI with current levels of safety and weak technical understanding presents an unacceptable risk**.

Regulation of AI

Governments all over the world are in the process of working out how to create clear, enforceable laws, often with the help of international organizations including the UN (and UNESCO specifically), OECD, World Economic Forum, and the Global Partnership on AI.

My July 25th written testimony lists several important proposals for regulation suggested by AI policy experts, so I will not repeat those here. I will focus on one basic principle: **the onus should be on developers to demonstrate rigorously to regulators that their systems are safe, before they can be released**. This is how we regulate nuclear power, aviation, and medicines.

While "safe" is currently too general and ill-defined a requirement, we can at minimum require proofs that systems will not exhibit certain obviously unacceptable behaviors, such as

- self-replication and cyberinfiltration of other computer systems;
- divulging classified information relevant to national security;
- defamation of real individuals.

Systems crossing these "red lines" should immediately be terminated and removed from the market, possibly with sanctions (e.g., fines) applied to the provider. In the case of open-source systems, this would apply to every copy, implying that copies must be self-registering and must include regulator-activated termination code.

Satisfying these requirements does not mean that an AI system is incapable of harm. They are necessary but not sufficient conditions for safety. An important side-effect of such rules would be to ensure that developers carry out further research on making AI systems predictable and controllable. This will contribute significantly to the long-term goal of making AI systems provably safe and beneficial.

Every state has a clear interest that AI systems remain safe and entirely under human control. Therefore, agreement should be possible, just as it has been in areas such as CFCs and nuclear safety, problems notwithstanding. An international coordinating body seems essential, as well as a coordinated international effort on AI safety research. Directions for research are listed in my July 25th testimony, including elements of a hardware-based proposal to ensure that unsafe AI systems cannot be deployed, either accidentally or deliberately.