Gradient Institute Ltd.
Level 2 Merewether Building H04
Cnr City Rd & Butlin Ave
The University of Sydney NSW 2006
https://gradientinstitute.org

Select Committee on Adopting Artificial Intelligence (AI)
Gradient Institute response 2 to question from Committee

Author: Bill Simpson-Young (Chief Executive)
Date: 17 June 2024

Please find below additional material in response to a question asked by the Committee at the Senate Hearing.

Committee question

Senator SHOEBRIDGE: If you were looking model by model and determining whether or not they were safe or the risks were acceptable, how would that work in practice?

Answer

With regard to frontier AI models, the problem of how to test these for safety is still an ongoing area of research and development. There is much more work to be done to address AI safety and some of this could be undertaken in Australia.

For example, some of the areas of frontier AI model risk for which AI safety testing methods and tools could be developed in Australia and which are in line with the Bletchley Declaration include the following (note: the below text is an extract from a January 2024 Gradient Institute document that proposes a Centre for AI Safety Testing (CAST) in Australia hosted by Gradient Institute - the full text is available on request from Gradient Institute):

—

The work of a Centre for AI Safety Testing (CAST) would entail technical research aimed at developing methods and tools for detecting and evaluating dangerous capabilities in AI systems, namely those falling under the categories of highest public concern expressed in the declaration: *misuse* and *misalignment* (or "control issues", in the language of the declaration). These methods and tools would be applied to test AI models and systems for government and business and, proactively, to test available open source models. Specifically, and also in line with the declaration, CAST would focus on the following areas of risk:

- ***AI-enhanced cybersecurity risks.*** One of the key areas of concern for AI misuse prioritised in the Bletchley declaration, and a major national security issue, is cybersecurity. That's well justified. Open-source LLMs can be easily reconfigured to override safety protections put in place by the original developer.[1] Safeguards of proprietary LLMs are routinely broken by experts, hackers, or technically savvy individuals, leading to these models behaving in hazardous ways.[2] Moreover, LLMs can code and make coders substantially more productive. Hackers can create AI hackers to help them. This can make it easier for bad actors to misuse LLMs for the purpose of unleashing devastating cyber-attacks. An example of a major concern amongst experts relates to the vulnerability of the power grid. An AI-enhanced attack successful at bringing down the power grid could be a major catastrophe. This could lead to widespread power outages, collapse of the healthcare system, breakdown of communication networks, interruption of the water supply, and widespread economic upheaval through the impairment of normal business operations and financial transactions. Clearly, developing effective testing protocols and tools to assess the risk exposure of an LLM to facilitate cyberthreats is paramount to ensure organisations only use LLMs with an acceptable cyber risk profile. This would be a major focus area of CAST.
- ***AI-enhanced biosecurity risks.*** Another major area of threat to national security, also prioritised in the Bletchley declaration, is biosecurity. Large Language Models (LLMs) are already capable of facilitating the synthesis of biological and chemical weapons,[3] as well as pandemic-class agents.[4] This raises essential security concerns for organisations of all types—whether they develop or use large language models (LLMs), including those proprietary, open-source, or developed in-house, across sectors such as technology, finance, telecommunications, government, and beyond. Especially when compounded with the cybersecurity risk, how can these entities ensure that their systems are safeguarded against adversarial attacks and unauthorised access, whether by internal employees or external parties, to exfiltrate from the model information that is not only "sensitive" but potentially catastrophic? Clearly, it makes sense to invest in developing tools to detect in an LLM the very existence of the capability to help facilitate the creation of dangerous biological or chemical compounds. This would be another major focus area of CAST.

[1] https://arxiv.org/abs/2311.00117
[2] https://arxiv.org/abs/2307.15043
[3] https://arxiv.org/abs/2304.05332
[4] https://arxiv.org/abs/2306.03809

Copyright © Gradient Institute 2024

- ***AI-enhanced disinformation, adversarial deception and manipulation***. The Bletchley declaration also provides great emphasis on the risk of disinformation, as well as the closely related issues of (adversarial) deception and manipulation. These risks are known to be key vectors of misuse of AI technology that can potentially threaten individual or public safety.
    - *Disinformation* is a form of deception that uses false information to target a large group of people or the general public. Generative AI can be used today to create false or misleading information to produce influence campaigns for a range of purposes, such as undermining trust in institutions, promoting ideologies, creating divisions and conflict, manipulating the market, sabotaging government policy (such as public health efforts), gaining geopolitical advantage, or distracting attention from other issues.[5]
    - *Adversarial deception* includes disinformation but can take other forms. It is the act of intentionally influencing people's beliefs away from the truth for personal gain. Generative AI and LLMs supercharge the potential for a wide range of deceptive practices, including creation of deepfakes, automation of phishing and spear phishing attacks, impersonation, identity fraud, spoofing, ad fraud and sales scams.[6]
    - *Manipulation* is the act of skillfully influencing someone's behaviour for personal gain, often concealing true intentions and methods, and possibly involving deception. There is great concern amongst experts that LLMs be used for large-scale and effective manipulation of individuals.[7] It is known that AI companions powered by LLMs are capable of generating profound emotional influence on people.[8] LLMs are known to be capable of being tailored for persuasive argumentation[9] and can be configured to behave manipulatively.[10]

Effectively *testing* whether a particular LLM is capable of or prone to being used for such purposes is key, as that's the basis for corrective and mitigative actions – technical, legal or otherwise. This is a significant challenge today. For instance, LLMs explicitly trained to deceive not only can do so effectively, but, concerningly, state-of-the-art methods to detect

[5] https://arxiv.org/abs/2301.04246

[6] https://arxiv.org/abs/2310.05189

[7] https://arxiv.org/abs/2306.11748

[8] https://www.abc.net.au/news/science/2023-03-01/replika-users-fell-in-love-with-their-ai-chatbot-companion/102028196

[9] https://arthurspirling.org/documents/llm.pdf

[10] In an ongoing technical project exploring the manipulation risks of LLMs, Gradient Institute has successfully configured a proprietary advanced large language model accessible through an API to behave manipulatively (towards another large language model).

and eliminate such deceptive capabilities have been shown to be ineffective.[11] CAST would focus on research to develop new methods for detecting and assessing these risks in LLMs, and develop tools implementing such methods to enable actual testing against these safety risks.

- ***Emergent deception and self-replication***. The risks discussed above fall into the class of *misuse*: the bad that comes from AI either following human intentions or being negligently handled by humans. The Bletchley declaration also puts emphasis on addressing risks of *misalignment,* or "issues of control". Absent bad intentions and negligence, things can still go wrong because it's not yet known, scientifically, how to control everything a frontier AI system does.[12] CAST would also devote attention to testing dangerous capabilities that may arise as a result of misaligned AI systems, as opposed to misuse by humans. Specifically, we plan to initially focus on two categories of misalignment risk:
  - *Emergent deception.* There is empirical evidence that deceptive capabilities have emerged in LLMs.[13] There is growing theoretical evidence that, as more powerful LLMs are created, more advanced deception capabilities are likely to emerge. [14] [15] [16] [17] This is of significant concern, since advanced forms of deception may include effective concealing of dangerous capabilities (including deception), thus directly threatening the very principle of AI Safety testing. Hence increasing attention in the AI Safety research community is being devoted to this problem. CAST would develop new research in emergent deception, aiming at developing testing protocols to detect the presence of such capabilities as well as assess the risk they pose to safety.
  - *Self-replication.* Another indicator that an AI model may be at risk of evading human control is if it develops the capability to create copies of itself. If an equivalent to an 'AI host cell' emerges through the training of an advanced AI model, it would serve as a significant alarm, indicating a potential loss of control. Self-replication is explicitly listed as one of the criteria against which the major AI companies

[11] https://www.nature.com/articles/d41586-024-00189-3

[12] This is often called the "alignment problem": how to effectively control AI systems (even those smarter than us). In its most general form, this is an open scientific problem and considered by experts an extremely difficult one to solve – if solvable at all. OpenAI in 2023 announced the creation of a new research unit focused on the mission of "solving super-alignment in four years", where "super-alignment" refers to alignment with AI systems much smarter than humans: https://openai.com/blog/introducing-superalignment

[13] https://arxiv.org/abs/2307.16513

[14] https://arxiv.org/abs/2209.00626

[15] https://onlinelibrary.wiley.com/doi/10.1002/aaai.12064

[16] https://arxiv.org/abs/2206.05862

[17] https://arxiv.org/abs/2206.13353

voluntarily agreed to testing their AI systems before release.[18] Indeed, OpenAI red-teamers tested GPT-4 against self-replication capabilities prior to its release (alongside numerous others).[19] CAST would dedicate a concentrated effort to produce novel methods and testing tools for detection of self-replication.

For further information

Committee Members are very welcome to contact Gradient Institute staff with further questions by emailing [redacted]

[18] https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[19] https://openai.com/global-affairs/our-approach-to-frontier-risk

Copyright © Gradient Institute 2024