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 Submission

Author: Bill Simpson-Young (Chief Executive)
Date: 20 May 2024

Note: I will present part of this submission to the Committee in a speech when I appear as a witness at the hearing on 20 May 2024. This submission is for providing details of my background to assist the Committee Members in understanding the scope of my knowledge with regard to the topic and for providing some core considerations related to AI that I believe are important for the inquiry.

About Gradient Institute

Gradient Institute is an independent, not-for-profit technology research institute that works to build safety, ethics, accountability and transparency into AI systems[1]. We perform research into improved AI algorithms, provide training and assessments for organisations which develop and deploy AI, and provide technical input into government AI policy development.

About the author

I am the Chief Executive of Gradient Institute and am a computer scientist and technologist. I first studied artificial intelligence (including neural networks) during the 1980s studying computer science at the University of Auckland and cognitive science at the University of NSW. I first worked in AI as a research assistant in machine learning at the University of Sydney in 1989, writing software that formed part of one of the first commercial machine learning engines.

Since then, I have spent my career so far leading teams of researchers, software engineers and designers to develop many novel computer technologies working in the Australian research and development labs of global technology companies (Canon and Unisys) and in government-funded research institutions (CSIRO and NICTA). My teams and I have built new technology and products

[1] https://www.gradientinstitute.org/

in areas such as machine learning, computer vision, natural language processing, data privacy, spatial data systems, internet technologies and computational law. Many of these technologies and products have included AI in them. Some of these teams have spun out to form new Australian technology companies. I have also been on the executive teams of three leading Australian research and technology organisations (Canon's R&D lab CiSRA, NICTA and CSIRO's Data61).

In 2019, after seeing the emergence of AI systems that had not been built rigorously and with proper consideration of their impact on people and society, I co-founded Gradient Institute to focus specifically on improving the safety and ethics of AI.

I am on the Australian Government's Temporary AI Expert Group, the NSW Government's AI Review Committee, the ANU School of Computing Advisory Board and several government research grant review committees. While at CSIRO's Data61, I was their representative on the Australian Government's Deputy Secretaries Data Group. I also designed and taught a Masters course in digital innovation at the University of Sydney for 7 years covering topics such as technology life cycles, disruptive innovation, open innovation and organisational culture for innovation.

Considerations for the Committee

The following is a list of points that I believe the Committee should consider while undertaking its inquiry into and report on the opportunities and impacts for Australia arising out of the uptake of AI technologies:

1. **AI is not just one technology :** The term "AI" covers a wide range of technologies with widely different sets of capabilities, supporting different sets of opportunities and with different risk profiles. Forms of AI have been used successfully in Australia, and with the relevant risks effectively managed, for decades. Widely used applications in which some form of AI is used include risk models, email spam detection, fraud detection, mapping apps, inventory management and many others which citizens and businesses rely on daily.

   Implication: It is important that concerns related to some of the most recent forms of AI (including generative AI) do not lead to actions that unduly impact the use of forms of AI that are already widely, safely and responsibly used. Any actions should be clear about precisely what technologies, properties or applications are in scope rather than referring to "AI" in general.

2. **AI capabilities have advanced extraordinarily rapidly in recent years and this trend is likely to continue into the coming years:** For example, OpenAI's GPT-3.5 AI model performed at the 10th percentile on the US Uniform Bar Exam compared to humans undertaking the same exam. Approximately 6 months later, its successor GPT-4 performed at the 90th percentile.[2] Capabilities such as converting text to video or live, interruptible human-like conversation were not even imagined by most people until they were announced by OpenAI[3]. The key drivers of this trend—an exponential increase in computational power and dataset sizes—show no clear signs of slowing down, indicating that the trend is likely to continue into the coming years.

   Implication: When considering actions related to AI, one shouldn't consider AI as being only what we now see it as. There need to be actions that take into account that radical changes in capabilities are likely to persist into the future.

3. **New capabilities emerge in foundation models without being designed in – including potentially dangerous capabilities:** In foundation models ("machine learning models trained on very large amounts of data that can be adapted to a wide range of tasks"[4]), new capabilities (such as the ability to translate between languages, to summarise documents and multi-step reasoning) have emerged without planning as the models have become larger and have had greater computational effort used in training[5]. These include potentially dangerous capabilities, such as a model developing its own goals and hiding them from its developers,[6] the ability to aid in the design of pandemic-grade pathogens,[7] or the ability for autonomously hacking websites.[8] New capabilities may be very useful but some may also be dangerous without careful controls.

   Implication: Actions on AI need to take into account that, for models beyond the frontier of what is known to be safe, effort by organisations and governments needs to go into ensuring the safety of the model development process and the final model produced before the model is deployed.

---

[2] https://arxiv.org/pdf/2303.08774 (Note: there has been some criticism of the methodology and so the actual percentile figure (https://link.springer.com/article/10.1007/s10506-024-09396-9) but there was massive improvement in capability nonetheless.)
[3] Sora (https://openai.com/index/sora/) and GPT-4o (https://openai.com/index/hello-gpt-4o/)
[4] https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
[5] https://arxiv.org/abs/2206.07682
[6] https://arxiv.org/abs/2401.05566
[7] https://arxiv.org/abs/2310.18233
[8] https://arxiv.org/abs/2402.06664

4. **AI will very likely outperform humans on most cognitive tasks within the next few years:** When measured against particular benchmarks, today's foundation models can already perform at a higher level than humans on benchmarks for areas such as reading comprehension, language understanding and image recognition and are approaching human performance on areas such as solving maths problems and computer software generation (see Figure 1)[9]. Given the rapid advancement and emergent capabilities already mentioned, it is expected this will continue for most human cognitive tasks within this decade.[10]

   Implication: Actions on AI must consider that, in the next few years, AI models will likely be capable of performing most of the economically valuable cognitive tasks currently performed by humans.

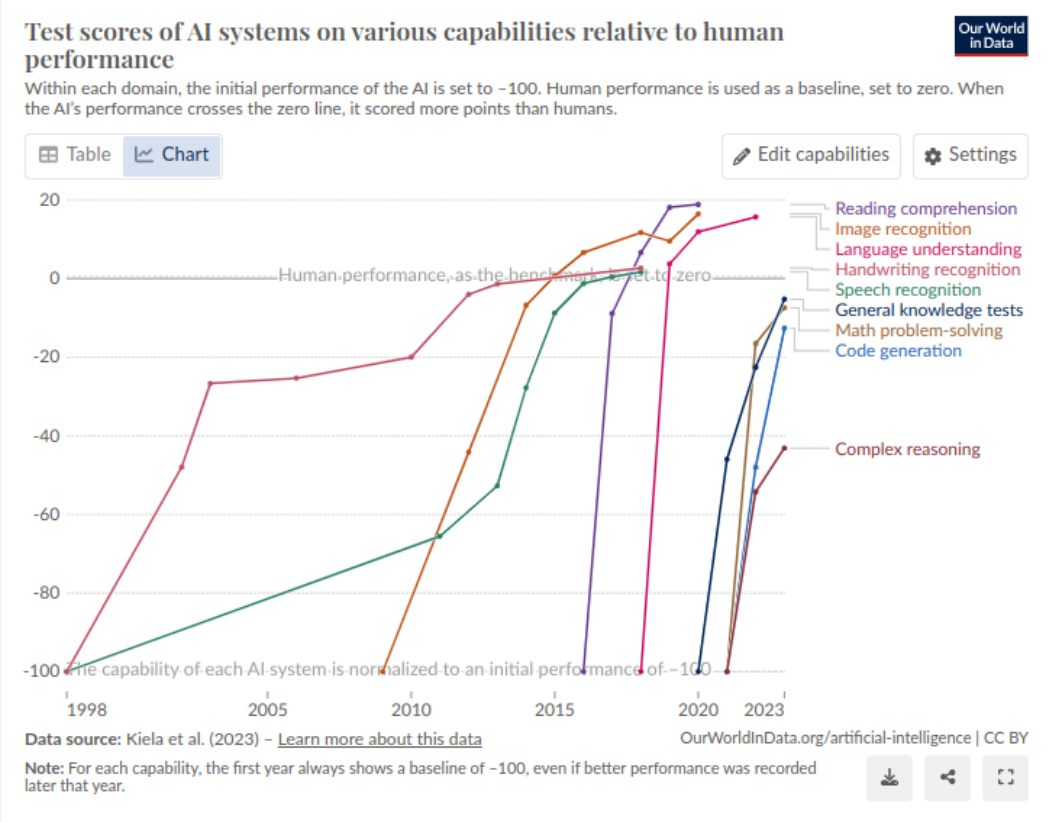

Figure 1: Advancing of cognitive capabilities toward and beyond human level (https://ourworldindata.org/artificial-intelligence)

[9] https://ourworldindata.org/artificial-intelligence

[10] The expected timeline for "Artificial general intelligence" (AGI), or AI as capable as humans in every cognitive domain, decreased by decades in just the past few years
https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/

5. **AI will lead to massive transformations on businesses, the economy and society:** Today's AI is already having a major impact on many industries and professions (such as software development, graphic design and copywriting) and, as AI models in the near future will outperform humans on many economically valuable cognitive tasks, future AI will transform society to a much greater level. This is likely to include major transformations in education and healthcare as well as in many other industries.

   Implication: Actions on AI should expect *paradigmatic* transformations of businesses, the economy and society, and not merely consider AI as a powerful tool making substantial changes to these.

6. **There is much innovation still to do in core AI research and technology development**: Despite the impressive capabilities of today's AI models and their many applications, there is much research and development work to be done to create the core AI technologies of the near future. For example, current foundation models are very expensive and energy-intensive to train and to operate and operate in a way that is not understood and with no built-in assurance of safety or reliability[11,12]. There are many opportunities for new architectures for AI models which are less resource-intensive and safer, new forms of large AI model that are useful for specific applications of AI (e.g. using forms of data other than language, images and audio such as DNA data, energy data etc), new techniques for training and testing AI models to ensure safer AI systems, and new tools and infrastructure that form the core of future AI systems.

   Implication: Australia should take action to ensure that it has a role in the new AI models and infrastructure that will be developed. This goes beyond AI research and goes beyond startups that merely *use* AI - it requires building an AI industry of research-backed core AI engineering capability. This is the type of AI engineering done at organisations such as OpenAI and in the R&D labs of large tech companies. Bootstrapping this requires government investment in areas such as talent attraction (e.g. a targeted AI research/engineer visa program), AI venture capital building (e.g. including overseas AI venture capital attraction and setting up a new focussed AI venture fund on core AI technology and not just startups using AI), an AI safety centre, etc.

[11] https://arxiv.org/html/2404.01157v1
[12] https://arxiv.org/pdf/2303.12712

7. **Many governments are investing in AI research, development and infrastructure**: Recognising the transformational impact to come from AI and the size of the opportunities, many countries have bet big on AI, including on research, development and infrastructure. Announcements made since January 2024 alone include large spending programs (mostly in the billions of dollars) by Canada, South Korea, Singapore, the UK and the US. Canada, for example, announced AU$2.7 billion investment in AI in April 2024. Such large investments are likely to lead to these countries (and others such as China who are already investing at a high rate) dominating in AI technology and possibly, through services available to companies and governments in other countries, controlling large parts of the working intelligence in the knowledge economy. It is also likely to lead to a global AI talent search by these countries.

   Implication: Australia should invest in AI research, development and infrastructure so that it is not left behind, so that it doesn't lose its AI talent and so that it avoids the possibility of effectively exporting large portions of its knowledge economy to foreign AI companies.

For further information

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