AIGS GSIA

ISED'S

AI COMPUTE ALLOCATION CONSULTATION

AIGS is pleased to present its recommendations in response to ISED's AI Compute Allocation Consultation

September
2024

Prepared By:
Wyatt Tessari l'Allie
Connor Fransoo
Kathrin Gardhouse

www.aigs.ca

September 6th, 2024

Jean-Philippe Lapointe
Director General
Innovation, Science and Economic Development
aicompute-calculia@ised-isde.gc.ca

Dear Mr. Lapointe,

We commend the Canadian government for turning its attention to AI compute and the pivotal role it can play in AI development, and thank you for the opportunity to provide comment on how to best invest the $2B announced in Budget 2024.

AI Governance & Safety Canada is a nonpartisan not-for-profit and a community of people across the country, working to ensure that advanced AI is safe and beneficial for all. We provided input to ISED on the Voluntary Code of Practice for Generative AI, participated in TBS's AI Strategy consultations, delivered extensive recommendations for Bill C-27, and most recently released our updated 2024 white paper *Governing AI: A Plan for Canada*.

Our submission is focused on the Compute Access Fund's potential implications for AI safety, and is structured as follows:

- The risks for the CAF that government will need to navigate
- The opportunity that the field of AI safety presents for the CAF
- Our three key recommendations for the CAF's success

Furthermore, we include as an annex our responses to select survey questions that are relevant to our expertise.

We remain available for any assistance that you require.

Sincerely,

Wyatt Tessari L'Allié
Founder & Executive Director
AI Governance & Safety Canada
contact@aigs.ca

Risk: the Compute Access Fund could exacerbate safety challenges

It is important to start by situating CAF in the larger AI context: with human intelligence staying the same, and artificial intelligence getting better by the day, we are headed into a world in which AI can outperform human beings in all domains. Building this level of AI is the explicit goal of many leading firms such as Google and Salesforce, and there are plausible scenarios in which it can be achieved in 2-10 years.

As AI capabilities continue to increase along this path, mitigating the associated risks becomes an increasingly difficult but essential task. Unfortunately, safety has until recently been an afterthought, and today it is vastly under-prioritised in industry compared to the size of the challenge. Moreover, public promises such as OpenAI announcing it would dedicate 20% of its resources to AI alignment have notoriously failed in practice.

The risk for the CAF is that it exacerbates this dynamic, pouring fuel on AI capabilities without adequately supporting safety, and tarnishes Canada's reputation in responsible AI. It is entirely possible that an AI model trained on government-funded compute causes a high-profile accident or societal harm, which could lead to public backlash and subsequent loss of funding for positive AI development.

Opportunity: Compute needs in AI safety

Fortunately, the global underinvestment in safety also creates a unique opportunity for the CAF to benefit society, research, and industry, and position Canada as a leader. AI safety covers a broad range of topics, including interpretability, alignment, model evaluation and monitoring, capability prediction, major accident mitigation, cybersecurity of powerful models, and more. Its success relies on availability of top talent, access to the latest models, and sufficient compute resources. We highlight here compute needs in two key domains:

Alignment

The work of guiding AI models toward human goals and values has recently seen a substantial growth in required compute resources as models have developed the ability to perform more complex, general-purpose behaviours. OpenAI's InstructGPT, an early example of a large-scale model fine-tuned for alignment, used just under "2% of GPT-3's pretraining compute and about 20,000 hours of human feedback" for alignment fine-tuning. As models become more complex, it is plausible that a greater percentage of the total training compute will need to be dedicated toward alignment. Furthermore, it is very likely that the total amount of compute used for training models will increase; EpochAI found that "the amortised hardware and energy cost for the final training run of frontier models has grown rapidly, at a rate of 2.4x per year since 2016" and that if trends continue, we will see billion-dollar training runs by 2027. Based on these trends and historical figures, we can estimate that alignment costs for a single frontier model will be in the tens or even hundreds of millions of dollars in the near future. Unfortunately, as described by AI safety researcher Jan Leike, companies are disincentivized from putting much effort toward aligning models, as high development and compute costs,

potential model performance degradation, and deployment time delays create a “commercial opportunity cost if you have customers who’d be willing to pay to use the unaligned model.”

Interpretability

While large-scale AI models have traditionally been seen as “black boxes”, raising concerns in fields like healthcare and criminal justice where explanation and justification of results are considered important, recent research in understanding how and why models produce the outputs they do, also called interpretability, has made some progress. Anthropic recently published one of the first studies on extracting interpretable, meaningful features from a medium-sized production model. However, they noted that they likely only extracted a small fraction of the total number of interpretable features and used a substantial amount of compute to do so, stating that to extract all features, they would “need to use much more compute than the total compute needed to train the underlying models”. While algorithmic improvements will clearly be needed to make true interpretability more practical, it is evident that there is an immediate need for compute resources in this sub-field of AI safety.

These two domains are only the tip of the iceberg in terms of compute needs, but they illustrate how the CAF can be used to significantly advance AI safety.

Recommendations

With developers under financial pressure to boost AI capabilities and cut corners on safety, government is in a unique position to alleviate this dynamic by allocating the CAF towards safety and ensuring AI remains beneficial to Canadians. Doing so will not only limit the risk of dangerous models being trained on government compute, it will enable the work of the new AI Safety Institute, retain and attract top talent to the field, play to our advantage in foundational research, and make Canada a global leader in this increasingly important field.

Recommendation	*Rationale*
1) Provide researchers and startups free compute for safety work *Target: minimum 20% of total CAF allocation to ensure sufficient impact*	By removing the compute costs for safety, it 1) Incentivises the work, 2) Allows government to track its progress, 3) Frees up capital for growth, hiring 4) Makes Canada a natural hub for research
2) Ensure all compute usage can be easily monitored and audited *Set up automated reporting to reduce burden on recipients*	Tracking how the compute is used is essential to ensure it is being allocated effectively, and not used to train harmful models.
3) Work with the AI Safety Institute, academia, and industry to further investigate and monitor safety needs	With the field rapidly evolving, continuously identifying needs and adjusting course is essential for the CAF safety investments to remain effective.

ANNEX - Responses to select survey questions

Note: AIGS Canada does not perform any activity requiring computing infrastructure. The following answers are provided to highlight how the Canadian AI Sovereign Compute Strategy and Compute Access Fund could positively impact the field of AI safety. They are based off our conversations with Canadian AI safety experts and other informed members of our community.

What is the benefit to you of having computing infrastructure that is Canadian-owned and controlled?

First, Canadian-owned and controlled computing infrastructure will reduce vulnerability in case of trade wars, supply chain issues, or foreign regulations on existing compute platforms. This will help all AI development in Canada, including safety work.

Second, it would mean the hardware can be identified, counted, tracked, and analysed, enabling government to understand the uses and development of AI in Canada, including the specific AI workloads being run. While this comes with potential privacy-related issues, there are technical solutions under development, such as “Cryptographic mechanisms on AI chips (that) could allow AI developers to securely log their workloads, which they could subsequently present to inspectors to attest their workloads” (Sastry, et al.). These solutions should be researched further to improve the monitoring of potentially dangerous AI training runs being conducted on CAF resources. These kinds of compute governance are especially valuable in the near term, before the development of distributed training methods that could limit our ability to monitor and make sense of training runs that only partly use Canadian hardware.

The third main benefit of Canadian-owned compute infrastructure is a potential increase in the productivity of safety researchers. Those whom we talked to said it is not uncommon to experience permission-related delays when trying to work with sensitive organisational data on foreign compute resources. Better hardware access will accelerate safety work.

How can we leverage this investment to both retain and attract AI talent in Canada?

During our conversations with AI safety researchers, the two main issues preventing the retention and attraction of AI safety talent in Canada were 1) low salaries, and 2) a lack of opportunity to conduct cutting-edge research in the most impactful safety topics. While the CAF or Sovereign Compute programs are unlikely to impact salaries directly, these investments would make AI safety research easier to conduct in Canada, provided that it dedicates compute resources for this field.

Decreasing the barriers to conducting productive research is likely to increase the quantity and quality of research being conducted, including in currently underexplored subfields

of AI safety. This increase in the breadth and depth of safety research would encourage researchers to come to or remain at our institutions, which would help develop, retain, and attract new AI safety talent here, and make Canada a global leader.

Are there any considerations that we have missed or elements we should explore further when addressing this topic?

For the CAF safety allocation to be effective, it is important to keep this category broad and not overly restricted. Breadth is important as AI safety research will need to cover a wide range of risks from both current and upcoming AI capabilities. And as with most scientific research, it's not always clear in advance which topics will turn out to be the most impactful.

Furthermore, for safety to keep pace with AI capabilities, it is important that research directions can pivot without facing overly restrictive rules. Striking this balance of prioritising important safety research without overly restricting its scope will be challenging. We therefore recommend taking an agile approach involving close monitoring of compute usage and ongoing consultations with stakeholders to ensure the CAF safety investments remain effective over time.