Comments on the interim report of the National Artificial Intelligence Research Resource Task Force

June 30, 2022

Lennart Heim
Research Scholar
Centre for the Governance of AI
lennart.heim@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

About the Centre for the Governance of AI (GovAI)

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK, with a US -presence. It was founded in 2018, initially as part of the Future of Humanity Institute at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI. More information at governance.ai.

Our comments

We welcome the opportunity to submit comments on the interim report of the National AI Research Resource's task force and look forward to future opportunities to input on the NAIRR. We offer the following submission for your consideration.

Key Recommendations

We recommend that the NAIRR:

Provides researchers with access to pre-trained models by

1. providing infrastructure that enables API-based research on large pre-trained models and guards against misuse
2. allowing researchers to use their NAIRR compute budget to do research on models accessed through an API
3. exploring ways to incentivize technology companies, academic researchers, and government agencies to provide structured access to large pre-trained models through the API

Addresses the risks stemming from AI models developed with NAIRR resources by

4. implementing a tiered access approach to compute provision, where access to larger amounts of compute comes with additional review requirements

Recommendations on topic (c)

(c) NAIRR resource elements and capabilities. Including data, government datasets, compute resources, testbeds, user interface, and educational tools and services. (Chapter 4 of the report)

Also [available on the GovAI blog](#).

Compute funds and pre-trained models

One of the key trends in AI research over the last decade is its growing need for computational resources. Since 2012, the compute required to train state-of-the-art (SOTA) AI models has been doubling roughly every six months[1]. Private AI labs are producing an increasing share of these high-compute SOTA AI models[2], leading many to worry about a growing compute divide between academia and the private sector[3]. Partly in response to these concerns, there have been calls for the creation of a National AI Research Resource (NAIRR)[4]. The NAIRR would help provide academic researchers with access to compute, by either operating its own compute clusters or distributing credits that can be used to buy compute from other providers[5]. It would also further support academic researchers by granting them access to data, including certain government-held datasets.

We argue that for the NAIRR to meet its goal of supporting non-commercial AI research[6], its design must take into account what we predict will be another closely related trend in AI R&D: an increasing reliance on large pre-trained models, accessed through application programming interfaces (APIs). Large pre-trained models are AI models that require vast amounts of compute to create and that can often be adapted for a wide array of applications. The most widely applicable of these pre-trained models have recently been called foundation models[7], because they can serve as a "foundation" for the development of many other models. Due to commercial considerations and concerns about misuse[8], we predict that private actors will become increasingly hesitant to allow others to download

[1] Sevilla et al., 2022
[2] According to Sevilla et al., 2022, every AI system that has set a new record for compute consumption since 2016 has been produced by a private lab.
[3] Ahmed & Wahed 2020; Ganguli et al. 2022
[4] Etchemendy & Li 2020
[5] Ho et al. 2021
[6] Ho et al. 2021
[7] Bommasani et al. 2021
[8] Brundage et al. 2018

copies of these models. We instead expect these models to be accessible primarily through APIs, which allow people to use or study models that are hosted by other actors. While academic researchers need access to compute and large datasets, we argue that they will also increasingly require API access to large pre-trained models. (Lohn & Musser have made similar claims.[9]) The NAIRR could facilitate such access by setting up infrastructure for hosting and accessing large pre-trained models and inviting developers of large pre-trained models (across academia, industry, and government) to make their models available through the system. At the same time, they could allow academics to use NAIRR compute resources or credits to work with these models.

The NAIRR has an opportunity, here, to ensure that academic researchers will be able to learn from and build upon some of the world's most advanced AI models. Importantly, by introducing an API, the NAIRR could provide structured access[10] to the pre-trained models so as to reduce any risks they might pose, while still ensuring easy access for research use. API access can allow outside researchers to understand and audit these models, for instance identifying security vulnerabilities or biases, without also making it easy for others to repurpose and misuse them.

Concretely, we recommend that the NAIRR:

1. provides infrastructure that enables API-based research on large pre-trained models and guards against misuse;

2. allows researchers to use their NAIRR compute budget to do research on models accessed through an API; and

3. explores ways to incentivize technology companies, academic researchers, and government agencies to provide structured access to large pre-trained models through the API.

Signs of a trend

We predict that an increasing portion of important AI research and development will make use of large pre-trained models that are accessible only through APIs. In this paradigm, pre-trained models would play a central role in the AI ecosystem. A large portion of SOTA models would be developed by fine-tuning[11] and otherwise adapting these models to particular tasks. Commercial considerations and misuse concerns would also frequently prevent developers from granting others access to their pre-trained models, except through APIs. Though we are still far from being in this paradigm, there are some early indications of a trend.

---

[9] Lohn & Musser 2022
[10] Shevlane 2022
[11] Fine-tuning describes the process of improving the performance of a pre-trained model on a specific task by training it on a task-related dataset.

Particularly in the domain of natural language processing, academic research is beginning to build upon pre-trained models such as T5, BERT, and GPT-3.[12] At one of the leading natural language processing conferences in 2021, EMNLP[13], a number of papers were published that investigated[14] and evaluated[15] existing pre-trained models. Some of the most relevant models are accessible only or primarily through APIs. The OpenAI API for GPT-3, announced in June 2020[16], has been used in dozens of research papers[17], for example investigating the model's bias[18], its capabilities[19], and its potential to accelerate AI research by automating data annotation[20]. Furthermore, Hugging Face's API interface has been used to investigate COVID-19 misinformation[21] and to design a Turing test benchmark for language models[22].

At the same time, in the commercial domain, applications of AI increasingly rely on pre-trained models that are accessed through APIs. Amazon Web Services, Microsoft Azure, Google Cloud[23], and other cloud providers now offer their customers access to pre-trained AI systems for visual recognition, natural language processing (NLP), speech-to-text, and more. OpenAI reported that its API for its pre-trained language model GPT-3 generated an average of 4.5 billion words per day[24] as of March 2021, primarily for commercial applications.

Five underlying factors in the AI field explain why we might expect a trend towards academic research that relies on large pre-trained models that are only accessible through APIs:

- Training SOTA models from scratch requires large amounts of compute, precluding access for actors with smaller budgets. For instance, PaLM[25] – a new SOTA NLP model from Google Research – is estimated to have cost between $9 and $23M to train.[26] The training compute cost of developing the next SOTA NLP model will likely be even greater.

- In comparison, conducting research on pre-trained models typically requires small compute budgets. For instance, we estimate that a recent paper investigating

[12] Raffel et al. 2019; Devlin et al. 2018; Brown et al. 2020
[13] EMNLP Conference 2021
[14] Wolfe & Clasikan 2021
[15] Elazar et al. 2021
[16] OpenAI 2020
[17] Google Scholar search
[18] McGuffie and Newhouse 2020
[19] Kohler and Daniel 2021
[20] Wang et al. 2021
[21] Wahle et al. 2021
[22] Uchendo et al. 2021
[23] Google 2022, AWS 2022, Microsoft Azure 2022
[24] OpenAI 2020
[25] Chowdhery et al. 2022
[26] Heim 2022

anti-muslim bias in GPT-3[27] likely required less than $100 of compute.[28] Developing new SOTA models by fine-tuning or otherwise adapting "foundation models" will also typically be dramatically cheaper than developing these models from scratch.

- The developers of large pre-trained models are likely to have strong incentives not to distribute these models to others, as this would make it both more difficult to monetize the models and more difficult to prevent misuse.

- Given the right infrastructure, it is significantly easier for researchers to use a pre-trained model that is accessed through an API than it is for them to implement the model themselves. This would enable user-friendly and secure access to the NAIRR (which is discussed in recommendation 4-20 of the interim report[29]). Implementing large models, even for research purposes, can require significant engineering talent, expertise, and computing infrastructure. Academics and students often lack these resources.

- Academics may increasingly aim their research at understanding and scrutinizing models, as this is important scientific work and plays to academia's comparative advantage.

We discuss these factors in detail in our blog post.

How the NAIRR could provide access to pre-trained models

We offer a sketch of how the NAIRR could provide access to pre-trained models in addition to data and compute, illustrated in the figure below. First, it would create a platform for hosting and accessing pre-trained models via an API. The platform should be flexible enough to allow researchers to run a wide range of experiments on a range of models. It should be capable of supporting fine-tuning, interpretability research, and easy comparison of outputs from multiple models. The API should allow researchers to interface with both models hosted by the NAIRR itself and models hosted by other developers, who may often prefer to retain greater control over their models.

Second, researchers would be allowed to use their NAIRR compute budgets to run inferences on the models. We recommend that researchers be allowed to use their budgets for this purpose even if the model is hosted by an organization other than the NAIRR.

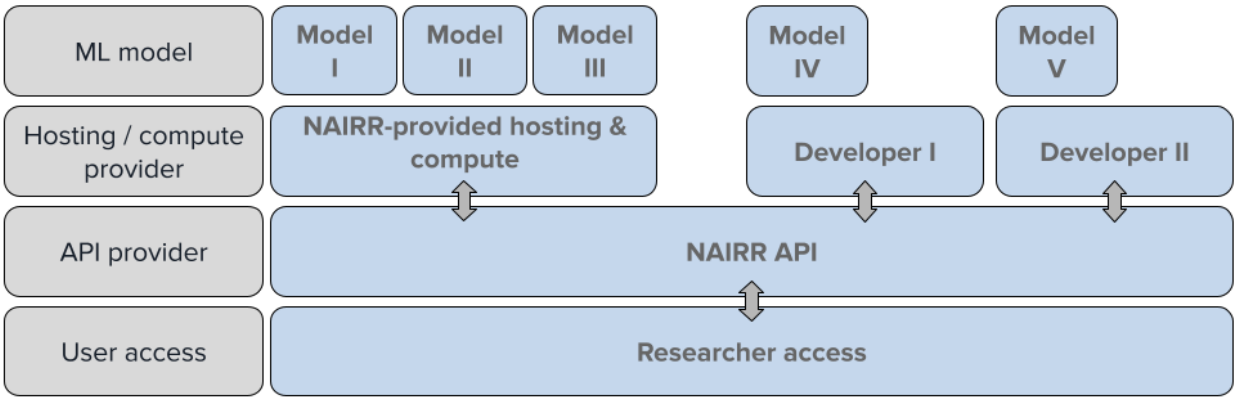
---

[27] Abid et al. 2021

[28] The authors probably used less than 10,000 prompts of around 20 tokens and received 10,000 outputs of around 20 tokens. This sums up to a total cost of around $24 via the OpenAI Davinci API ($0.06 per 1,000 tokens). This would be cheaper if using a less powerful version of GPT-3 or when the inference is self-hosted.

[29] "*Recommendation 4-20: To help realize its vision, the NAIRR must provide secure and user-friendly access to integrated services, resources, data, and training materials.*"

In recommendation 4-13 of the interim report, "three levels" of the NAIRR compute resources are suggested.[30] The proposed API would be part of the third and highest level — providing access to pre-trained models for a wide range of users.



An illustration of how the NAIRR could provide API access to large pre-trained models.

The biggest challenge will likely be securing access to pre-trained models from developers across industry, academia, and government. In some cases, developers might be motivated to provide access by a desire to contribute to scientific progress, the prospect of external actors finding issues and ways to improve the model, or a belief that it might improve the organization's reputation. The NAIRR could also create an expectation that models trained using NAIRR compute should be accessible through the platform. Access to particularly high-stakes government models in need of outside scrutiny could also potentially be mandated. Additionally, the NAIRR could consider incentivizing government agencies to provide API access to some of their more impactful models in exchange for access to compute resources or data (similar to a Stanford HAI proposal regarding data access[31]).

Encouraging private actors to make their models accessible through the platform may be especially difficult. In some cases, companies may provide model access as a means to build trust with their consumers. They may recognize that the public will be far more trusting of claims concerning the safety, fairness, or positive impacts of their AI systems if these claims are vetted by outside researchers. For example, Facebook and Twitter have recently created APIs that allow outside researchers to scrutinize company data in a privacy-preserving manner.[32] Further, the NAIRR could consider offering compensation to developers for making their models available via the API. Developers may also be particularly concerned about risks to intellectual property, something that can be assuaged by the NAIRR upholding high cybersecurity standards.

---

[30] "*Recommendation 4-13: Software leveraged for NAIRR compute resources should span three "levels" to support a broad user base.*"
[31] Ho et al. 2021
[32] TechCrunch 2021; Twitter 2022

Crucially, the API should also be designed to thwart model misuse, while still ensuring easy access for research use. Multi-purpose models trained with NAIRR resources could be used maliciously, for instance by criminals, propagators of misinformation, or autocratic governments around the world. Large language models could, for example, significantly reduce the cost of large-scale misinformation campaigns[33]. The NAIRR should take measures to avoid models trained with publicly funded compute being put to such uses. Misuse could be reduced by introducing a tiered access approach, as suggested in the Stanford HAI report[34] for datasets hosted on the NAIRR. For instance, researchers might get easy access to most models but need to apply for access to models with high misuse potential. Further restrictions could then be placed on the queries or modifications that researchers are allowed to make to certain models. In addition, API usage should be monitored for suspicious activity (e.g. the generation of large amounts of political content).

Helping academic researchers share their models

An appropriately designed API could also solve a challenge the NAIRR will face as it provides compute and data for the training of large-scale models: academic researchers will likely want to share and build on models developed with NAIRR resources. At the same time, open-sourcing the models may come with the risk of misuse in some cases. By building an API and agreeing to host models itself, the NAIRR can address this problem: it can make it easy for researchers to share their models in a way that is responsive to misuse concerns.

Academics are significantly more likely to voluntarily make their models available via the API than private developers of SOTA models with a profit motive. As such, the NAIRR could start by focusing on providing infrastructure for academic researchers to share their models with each other, thereby building a proof-of-concept, and later introducing additional measures to secure access to models produced in industry and across government. Eventually, the NAIRR could set a standard — enabling a vibrant and growing AI ecosystem, as proposed in recommendation 4-22[35], while maintaining critical security needs.

Conclusion

By building API infrastructure to support access to large pre-trained models, the NAIRR could produce a number of benefits. First, it could help academics to scrutinize and understand the most capable and socially impactful AI models. Second, it could cost-effectively grant researchers and students the ability to work on frontier models. Third, it could help researchers to share and build upon each other's models while also avoiding risks of misuse. Concretely, we recommend that the NAIRR:

---

[33] Weidinger et al. 2021; Buchanan et al. 2021
[34] Ho et al. 2021
[35] "*Recommendation 4 -22: The NAIRR should embrace standards, including de facto standards, and best-of-breed open-source solutions whenever possible to ensure a vibrant, growing AI ecosystem*."

1. provides infrastructure that enables API-based research on large pre-trained models and guards against misuse;

2. allows researchers to use their NAIRR compute budget to do research on models accessed through an API; and

3. explores ways to incentivize technology companies, academic researchers, and government agencies to provide structured access to large pre-trained models through the API.

Recommendations on topic (d)

(d) System security and user access controls. (Chapter 5 of the report)

We recommend that the NAIRR task force implements a tiered access scheme to computational resources (in short *compute*) — similar to the recommendation for the access to sensitive and private data.[36] Since compute is a finite, rivalrous resource, the NAIRR will have to make difficult decisions about how it is allocated. Such decisions should be based on many factors, including scientific merit and practicability. Importantly, it should also be based on the extent to which the researchers adhere to responsible AI practices, e.g. foreseeing and preventing potential risks the model could impose. The more compute a project is allocated, we argue, the greater care should be taken by the NAIRR and the researchers to reduce risks and spread the benefits of the system.

Chapter 5 of the interim report outlines the security and user access to the NAIRR. We welcome and support the outlined recommendations for protecting sensitive and private data. Nonetheless, the recommendations do not sufficiently address the potential risks stemming from AI systems created with resources by the NAIRR. As many scholars have argued, AI systems can pose a variety of risks and should undergo an extended review process before their creation and potential publication.[37] As an example, systems that can read and write can substantially impact daily life,[38] and their surprising and unpredictable capabilities[39] warrant an extensive review and monitoring process. The NAIRR should facilitate and enforce responsible development and disclosure of these powerful AI systems. The NAIRR should become leader in the co-development of these guidelines and enforce them for research conducted using the NAIRR — helping to set the standards for responsible AI.[40]

---

[36] See Recommendation 4-9 (p.4.5) and 5-3 (p.5.3) of the interim report.
[37] Brundage et al. 2018 "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation."
[38] See a recent discussion by leading AI labs on "Best Practices for Deploying Language Models"
[39] Ganguli et al. 2022 "Predictability and Surprise in Large Generative Models"
[40] As outlined in the Executive Summary: "*The NAIRR can set the standard for responsible AI research through the design and implementation of its governance processes.*" (p.iii)

Why use compute as a proxy for the potential impact of a system?

The compute used to train an AI system is a particularly useful metric when considering what level of responsible AI practices should be demanded. Firstly, the performance of machine learning models tend to scale with compute.[41] State-of-the-art models across domains, such as PaLM, AlphaFold, GPT-3, have one thing in common: they use a large amounts of compute.[42] For example, it took more than 64 days across thousands of chips to train PaLM with an estimated cloud computing cost of $9M to $23M.[43] While the performance of a system also scales with the amount and quality of data,[44] there are no agreed-upon metrics of data quality that could be used for this purpose. Secondly, the performance of an AI system is a useful proxy of its potential impact, both positive and negative. The more capable the system, the more uses it can be put to, and the more important it is that it is developed and deployed responsibly.

Other metrics should also be considered. For example, the NAIRR could introduce stricter requirements for AI systems used in particularly high risk domains such as health care or biometric identification. However, it is often difficult to predict the downstream impacts of an AI system or research contribution.[45] An AI system developed for one use can often be put to others: an AI drug discovery tool could be repurposed to design biochemical weapons or other toxic substances.[46] Further, high-compute models trained today – and likely trained using NAIRR resources – tend to be general AI systems[47], where it is even more challenging to predict the uses or even capabilities.[48] As such, considering only, for example, the uses a model will be put to is not sufficient.

What could a compute-based tier of responsible AI practices look like in practice?

Researchers or projects receiving small amounts of compute could be subject to minimal or no responsible AI requirements. They could be required to submit a description of what they plan to use the resources for and sign an agreement saying that they will adhere to a NAIRR code of conduct. The NAIRR should also do spot checks to see if compute is being used for the intended purpose.

---

[41] Kaplan et al. 2020 "Scaling Laws for Neural Language Models"; Hofman et al. 2022 "Training Compute-Optimal Large Language Models"
[42] Sevilla et al. 2022 "Compute Trends Across Three Eras of Machine Learning"
[43] Chowdhery et al. 2022 "PaLM: Scaling Language Modeling with Pathways"; Heim 2022 "Estimating PaLM's training cost"
[44] Model size (number of parameters) and number of data samples are linear correlated with the amount of compute. However, it's indepdent of the quality of data.
[45] Prunkl et al. 2021 "Institutionalizing AI Ethics via Broader Impact Statements"
[46] Urbina et al. 2022 "Dual use of artificial-intelligence-powered drug discovery"
[47] Bommasani et al. 2021 "On the Opportunities and Risks of Foundation Models"
[48] Ganguli et al. 2022 "Predictability and Surprise in Large Generative Models"

At the higher end, a number of requirements could be imposed on the project. As a starting point, such developers of such models could be required to adhere to the forthcoming NIST AI Risk Management Framework[49], in addition to an extended review process and policies around the future publication and usage.

Importantly, a number of measures could be taken to ensure that potential risks from the system are identified and mitigated. Identifying such risks can be hard, as it is difficult to predict what tasks a general model will perform well at,[50] and because the eventual impacts of the system depends on how it gets incorporated into larger sociotechnical systems. This could be done by requiring external audits or red team exercises. It could also be done by giving initial access to a few dozen researchers and giving them a bias/safety bounty if a flaw in the model is identified.[51]

Potential risks from the system could be addressed in the development phase, ensuring that the models are sufficiently accurate, fair, robust, aligned, interpretable and the like. Some risks can be addressed via appropriate deployment strategies or "structured access".[52] For example, particularly general and capable models could be made available to a wide audience using an API, with monitoring and filters to prevent misuse.

Deciding on the thresholds for the responsible AI tiers will be a challenging task. A useful starting reference might be a fraction of the compute used for the final training run of state-of-the-art AI systems, likely measured in FLOPs. The final training run compute is commonly reported by researchers and scholars have tracked it over time.[53] While those reports only address the compute used for the final training run, not the complete development process, this can be used as a lower bound for the required compute for developing AI systems. Further, the compute tresholds should likely differ depending on the type of system or application domain, as e.g. SOTA models in protein folding require much less compute than those in natural language processing.

[49] NIST 2022 AI Risk Management Framework
[50] Ganguli et al. 2022 "Predictability and Surprise in Large Generative Models"
[51] For a description of these tools and more, see Brundage et al., 2021. "Toward Trustworthy AI Development"
[52] Shevlane 2022 "Structured access: A paradigm for safe AI deployment"
[53] Sevilla et al. 2022 "Compute Trends Across Three Eras of Machine Learning"; Amodei & Hernandez 2018 "AI and Compute"