PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

September 1, 2021

Via email: NAIRR-responses@nitrd.gov
White House Office of Science and Technology Policy and National Science Foundation
Wendy Wigen, NCO,
2415 Eisenhower Avenue,
Alexandria, VA 22314

RFI Response: National AI Research Resource

Thank you for the opportunity to respond to the National AI Research Resource (NAIRR) RFI. We are academic researchers associated with the Center for Information Technology Policy (CITP) at Princeton University[1] and write in support of the Task Force's aim to create equitable access to the infrastructure that fuels AI research and development.

In our response, we highlight the significance of supporting a research infrastructure that is designed to independently test the validity of the claims of AI performance. In particular, we draw attention to the widespread phenomenon of the industry peddling what we call "AI snake oil" — promoting an AI solution that cannot work as promised.[2] Relatedly, we highlight how AI-based scientific research is often plagued by overly optimistic claims about its results and suffers from reproducibility failures. We submit that the Task Force's implementation roadmap for the NAIRR must include establishing a public infrastructure that can critically evaluate AI performance claims. This infrastructure is vital to the goals of the Task Force of ensuring that AI research serves our shared democratic values.

[1] In keeping with Princeton's tradition of service, CITP's Technology Policy Clinic provides nonpartisan research, analysis, and commentary to policy makers, industry participants, journalists, and the public. This response is a product of that Clinic and reflects the independent views of the undersigned scholars.
[2] *How to Recognize AI Snake Oil*, Arvind Narayanan, Nov. 18, 2019, available at https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

1. We need a research infrastructure that critically evaluates AI-based performance claims and ensures that those tools are designed to serve societal values. *(Response to Question 3.)*

Recently, the industry has converged on a troubling and widespread practice that applies the label of AI to applications that do not and cannot work. We dub this phenomenon of using a veneer of AI to lend credibility to pseudoscience as *AI snake oil*. The proliferation of AI snake oil in such applications is a distinct issue from concerns around bias, but is a major contributor to the negative consequences that result.

AI-based research has led to undeniable genuine and rapid progress in many domains, but it is important to distinguish between the classes of problems where AI tools have been shown to be effective. For example, AI has made significant progress in aiding with perception tasks, but it has struggled to predict outcomes involving complex social phenomena. Applications that claim to predict social outcomes but in fact do not have any predictive power are unfair even if they are technically unbiased, since they mask the fact that they do not work as promised and end up perpetuating outcomes that are not well calibrated to the needs. This is especially true when they are deployed in determining important life outcomes.

As an example, consider the AI tools that are purportedly designed to automate hiring decisions. The main claim made by many companies producing these tools is that AI can analyze body language and personality traits from short videos of candidates and function as "algorithmic pre-employment assessments" to make hiring decisions easier. While it is generally understood by experts that these tools cannot work and are usually no better than random number generators, that has not stopped companies from riding the AI hype and being widely funded and adopted. Raghavan et al. highlight that 18 companies working on algorithmic hiring systems have collectively raised over $200 Million in funding over the last few years, though not all of these companies offer AI assessments of job candidates.[3]

[3] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." ACM Conference on Fairness, Accountability, and Transparency.

Similar claims prevail in a large number of applications where AI systems are claimed to predict social outcomes such as the likelihood of recidivism or identifying at-risk kids. But recent research shows that AI systems today are no better than simple rules at predicting social outcomes.[4] However, this does not stop companies from marketing AI-based systems that claim to solve these problems, and as a result industrial applications of AI that purportedly predict social outcomes are proliferating. This phenomenon has a further pernicious effect of fueling the hunger for personal data for these fundamentally dubious applications of AI and giving rise to "black box" algorithms that cannot be explained. Furthermore, these applications tend to distract attention from designing more effective interventions.

As a result, we see evaluating validity as a core component of ethical and responsible AI research and development. The Task Force could support such efforts by setting standards for and making tools available to independent researchers to validate claims of effectiveness. The NAIRR could also help create oversight mechanisms and support efforts to regulate AI tools that are known to not work.

2. There is a reproducibility crisis in scientific research that relies on AI and machine learning that the Task Force should address. *(Response to Question 3.)*

Scientific research suffers from a closely related problem to the industry's reliance on AI snake oil. Many studies that purport to rely on AI have results that are overly optimistic and lack reproducibility.[5] But there are challenges in creating the incentives for researchers to independently and rigorously examine scientific claims that the NAIRR can help us overcome.

Evaluating academic claims about machine learning is challenging. First, the code tends to be complex and lacks standardization, which makes it difficult to understand and replicate models. Second, there are subtle pitfalls for researchers who fail to differentiate between explanatory and predictive modeling. Third, the hype and overoptimism about commercial AI often spills over into machine learning research and obscures the findings.[6] All these, of course, are in addition to

[4] Matthew J. Salganik et al. 2020. "Measuring the predictability of life outcomes with a scientific mass collaboration." Proceedings of the National Academy of Sciences 117 (15).
[5] Sayash Kapoor and Arvind Narayanan. 2021. "(Ir)reproducible Machine Learning: A Case Study." Preprint available at reproducible.cs.princeton.edu.
[6] Joelle Pineau et al. 2020. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." arXiv preprint arXiv:2003.12206.

the pressures and publication biases present in all disciplines that have led to reproducibility crises.

Systematic reviews have started to identify reproducibility issues and overoptimistic results in many academic fields that are adopting machine learning methods (*see* Figure 1 below). But this is complex and expensive work. One estimate suggests that we spend over $28 billion a year on preclinical research in the United States that is not reproducible.[7] As machine learning methods spread across academic fields, focusing on the reproducibility of that research is critical to ensure its validity.

One of the major roadblocks to reproducibility research is that appropriate computing resources are difficult to secure. While researchers can rely on cloud services such as Amazon AWS, Google Cloud and Microsoft Azure for compute-intensive AI research, there are fewer resources available for those seeking to vet claims of performance. This problem has intensified with the shift of private firms undertaking research into new AI models. For example, natural language processing models routinely require large amounts of computational resources. But the cost of computational resources to replicate performance claims are often beyond the reach of independent researchers at research universities. This further makes reproducibility of research output by private companies inaccessible due to issues with data sharing and lack of access to computational infrastructure.

We recommend that the NAIRR prioritizes the support of systematic reviews of published research across fields adopting machine learning methods. For example, the NAIRR could establish and sustain a computational reproducibility infrastructure and serve as a reproducibility clearinghouse by setting up benchmark datasets for measuring progress.[8] This would lead to significant strides towards the aim of promoting transparent, effective, and responsible research.

[7] Leonard P. Freedman , Iain M. Cockburn, Timothy S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." PLoS Biology 13(6).
[8] Benjamin Haibe-Kains et al. 2020. "Transparency and reproducibility in artificial intelligence." Nature 586, E14–E16.

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Neuroimaging	Whelan et al.	2014	—	4	Incorrect train-test split
Autism Diagnostics	Bone et al.	2015	2	2	Biased evaluation; data leakage
Bioinformatics	Blagus et al.	2015	—	6	Data leakage
Nutrition research	Ivanescu et al.	2016	—	4	Incorrect train-test split
Text Mining	Olorisade et al.	2017	30	—	Multiple pitfalls
Clinical epidemiology	Christodoulou et al.	2019	71	48	Biased evaluation; data leakage
Recommender Systems	Dacrema et al.	2019	26	25	Weak baselines, don't share code or data
Toxicology	Alves et al.	2019	1	1	Multiple pitfalls
Computer security	Arp et al.	2020	30	30	Multiple pitfalls
Health care	McDermott et al.	2021	511	—	Multiple pitfalls
Medicine	Vandewiele et al.	2021	24	21	Incorrect train-test split, data leakage
Radiology	Roberts et al.	2021	62	62	Multiple pitfalls

Figure 1 [from Kapoor and Narayanan]: a list of systematic reviews that highlight overoptimism and irreproducibility in applied machine learning research across academic fields.

3. The NAIRR can promote effective data stewardship models for using datasets. *(Response to Question 2, Item D.)*

The creation of datasets has been pivotal in the development of AI applications. But there is an underexplored dark side to supporting the broad release of datasets without mechanisms of oversight or accountability for how that information can be used. The resulting harms include privacy risks and representational harms. The NAIRR can play a pivotal role in mitigating these harms by establishing and supporting appropriate data stewardship models.

Consider the challenge of "runaway datasets" as an example of a problem that the NAIRR might address. In the last few years, many datasets have been retracted due to ethical concerns. But our research has documented how, even

after retraction, these datasets can remain widely available and are used across the industry and in research labs.[9] This phenomenon has been dubbed the problem of "runaway datasets." Of course, the ethical issues that caused the researchers to retract the original dataset persists in AI applications that continue to use these datasets after retraction. This highlights the necessity of dealing with ethical issues throughout the lifecycle of the dataset instead of addressing ethical issues only when the dataset is released.

In particular, the existing ethical oversight mechanisms within academia such as IRBs (Institutional Review Boards) are poorly suited to deal with runaway datasets. "Human subjects research" has a narrow definition in the context of IRBs and thus many of the datasets and associated research that have caused ethical concern in machine learning would not fall under the purview of IRBs. Further, IRBs do not consider downstream harms during their appraisal of research projects.[10] This compounds issues with runaway datasets and exacerbates ethical concerns with the creation and use of datasets.

The NAIRR can address this gap by creating centralized data clearinghouses to regulate access to datasets. Such clearinghouses could include safeguards for monitoring ethical concerns through the lifecycle of the use of the datasets. The NAIRR could also create a framework for licensing datasets and machine learning models so researchers can control the intended and acceptable uses of their work. For example, we see significant confusion resulting from the use of unclear and non-standardized licenses in dataset releases. Finally, the NAIRR could establish mechanisms for exercising responsible data stewardship that can make decisions about the ethical uses of datasets at the time they are being created and while they are in use. While some research projects already follow such a procedure when releasing datasets, institutional support including providing funding towards data stewardship committees would help reduce the ethical risks of AI applications due to runaway datasets.[11]

* * *

[9] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. "Mitigating dataset harms requires stewardship: Lessons from 1000 papers." arXiv preprint arXiv:2108.02922.
[10] Jacob Metcalf. 2017. "The study has been approved by the IRB': Gayface AI, research hype and the pervasive data ethics gap." Pervade Team.
[11] Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J. Salganik. 2018. "Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge." Socius, 5.

We commend the Task Force's careful attention to these issues and welcome the opportunity to discuss any questions.

Respectfully submitted,

Sayash Kapoor
Graduate Student, Department of Computer Science

Mihir Kshirsagar
Technology Policy Clinic Lead, Center for Information Technology Policy

Arvind Narayanan
Associate Professor of Computer Science

Contact:

Website: https://citp.princeton.edu
Phone: 609-258-5306
Email: mihir@princeton.edu