March 4, 2022

Via email: AI-RFI@nitrd.gov
White House Office of Science and Technology Policy,
Faisal D'Souza, NCO,
2415 Eisenhower Avenue,
Alexandria, VA 22314

Response to Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan

Thank you for the opportunity to respond to the Request for Information to the Update of the National Artificial Intelligence Research and Development Strategic Plan ("Strategic Plan"). We are academic researchers associated with the Center for Information Technology Policy (CITP) at Princeton University,[1] Microsoft Research, and Cornell University, and write to provide suggestions for how the Strategic Plan can focus resources to address societal issues such as equity, especially in communities that have been traditionally underserved. We also discuss how AI R&D can support research that informs the intersection of AI R&D and its application with privacy and civil liberties.

1. Strategy 1: Sustaining long-term investments in fundamental AI research requires supporting research on its impact on equity.

The 2019 Update and the original Strategic Plan rightly emphasize the importance of sustaining long-term investments in fundamental AI research. One core area for support that the Strategic Plan highlights is investments to advance trust in AI systems, which includes requirements for robustness, fairness, explainability, and security. This area of research has only become more important to sustain as AI systems have become embedded in public life. But, we suggest, the Strategic Plan should also explicitly include a commitment to making investments in research that examines how AI systems can affect the equitable distribution of

[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.

resources. Specifically, there is a risk that without such a commitment, we make investments in AI research that can marginalize communities that are disadvantaged. Or, even in cases where there is no direct harm to a community, the research support focuses on classes of problems that benefit the already advantaged communities, rather than problems facing disadvantaged communities.

We recommend that the Strategic Plan outline a mechanism for a broader impact review when funding AI research. As we argue below (Section 2), existing institutional mechanisms for ethics review of research projects do not adequately identify downstream harms stemming from AI applications. When deciding where to invest resources, the government and its funding bodies should take into account not only the potential positive impacts of research, but the potential negative impacts as well. The Strategic Plan should include mechanisms that take advantage of the government's unique position to steer the research community away from research questions that pose obvious risks of downstream harm without any clear benefits, such as the many phrenology-like studies in computer vision that have generated recent controversy.[2]

Because AI research can sometimes result in rather general knowledge or techniques with a broad range of potential applications, it may be challenging to determine what kind of impact it might have. In fact, many AI research findings will have dual use: some applications of these findings may promise exciting benefits, while others would seem likely to cause harm. While it is worthwhile to weigh these costs and benefits, decisions about where to invest resources should also depend on distributional considerations: who are the people likely to suffer these costs and who are those who will enjoy the benefits? Research should not only have a positive broader impact; its benefits should be distributed equitably. In fact, even research that only seems to have a positive upside should be assessed with distributional concerns in mind to ensure that the benefits don't accrue primarily to those who are already advantaged in society. While there have been recent efforts to incorporate ethics review into the publishing processes of the AI research community,[3] [4] adding similar considerations to the Strategic Plan would help to highlight these concerns much earlier in the research process. Evaluating research proposals according to these broader impacts would help to ensure that

[2] Luke Start and Jevan Hutson. "Physiognomic Artificial Intelligence." Fordham Intellectual Property, Media & Entertainment Law Journal, 2021.
[3] Brent Hecht et al. "It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process". ACM Future of Computing Blog, 2018.
[4] Priyanka Nanayakkara, Jessica Hullman, Nicholas Diakopoulos. "Unpacking the Expressed Consequences of AI Research in Broader Impact Statements." AIES, 2021.

ethical and societal considerations are incorporated from the beginning of a research project, instead of remaining an afterthought.

2. Prioritize research on the downstream implications of AI research and applications under Strategy 3 of the Strategic Plan.

The Strategic Plan correctly focuses on supporting research that designs architectures for ethical AI. But, on privacy issues, ethical AI has sometimes been framed incorrectly as merely concerning the data collection and management process.[5] We suggest that a larger threat comes from the downstream impacts of AI applications such as face recognition,[6] workplace surveillance,[7] and behavioral advertising.[8]

The current Strategic Plan focuses on two notions of privacy: (i) ensuring the privacy of data collected for creating models via strict access controls, and (ii) ensuring the privacy of the data and information used to create models via differential privacy when the models are shared publicly. Both of these approaches are focused on the privacy of the people whose data has been collected to facilitate the research process, not the people to whom research findings might be applied. Take, for example, the potential impact of face recognition for detecting ethnic minorities.[9] Even if the researchers who developed such techniques had obtained approval from the IRB for their research plan, secured the informed consent of participants, applied strict access control to the data, and ensured that the model was differentially private, the resulting model could still be used without restriction for surveillance of entire populations,[10] especially as institutional mechanisms for ethics review such as IRBs do not consider downstream harms during their appraisal of research projects.[11]

While it is critically important to protect the privacy of the people whose data are being used in the research process, such protections do nothing to ensure

---

[5] Vinay Uday Prabhu and Abeba Birhane. “Large image datasets: A pyrrhic win for computer vision?” arXiv preprint arXiv:2006.16923, 2020.
[6] Antoaneta Roussi. “Resisting the rise of facial recognition.” Nature news feature, 2020.
[7] Kyle Wiggers. “Workplace surveillance algorithms need to be regulated before it’s too late.” VentureBeat, 2021.
[8] Charles Duhigg. “How Companies Learn Your Secrets.” New York Times, 2012.
[9] Richard Van Noorden. “The ethical questions that haunt facial-recognition research.” Nature News Feature, 2020.
[10] Solon Barocas and Helen Nissenbaum. “Big Data’s End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, Eds. Cambridge University Press, NY, 2014.
[11] Jacob Metcalf. 2017. “The study has been approved by the IRB': Gayface AI, research hype and the pervasive data ethics gap.” Pervade Team.

that the resulting discoveries do not threaten other people's privacy. Even if the data used for creating the models is stored privately, the models created using this data can still be used for privacy-breaching inferences. In fact, even if the data that was collected for training the AI model is later deleted, the models trained using this data can still be used for such inferences. And models that are differentially private are just as good at privacy-breaching inferences as those that are not differentially private.

The Strategic Plan must therefore grapple with the fact that AI applications are a powerful tool for privacy-breaching inferences—even when the underlying research has taken the privacy interests of research subjects into account. We recommend that the Strategic Plan include as a research priority supporting the development of alternative institutional mechanisms to detect and mitigate the potentially negative downstream effects of AI systems. In addition, we recommend that the Strategic Plan include provisions for funding research that would help us understand the impact of AI systems on communities, and how AI systems are used in practice. Such research can also provide a framework for informing decisions on which research questions and AI applications are too harmful to pursue and fund.

3. Prioritize systematic studies of reproducibility under Strategies 5 and 6.

Many studies that purport to rely on AI have results that are overly optimistic and lack reproducibility.[12] Indeed, we found 18 reviews across 15 scientific fields that find errors in a total of 304 papers that use ML-based science (*see* Figure 1 below). Given the adoption of ML methods across scientific fields, there is an urgent need to address reproducibility issues in ML-based science. But there are challenges in creating the incentives for researchers to independently and rigorously examine scientific claims that the Strategic Plan can help 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.[13] All these, of course, are in addition

[12] Sayash Kapoor and Arvind Narayanan. 2021. "(Ir)reproducible Machine Learning: A Case Study." Preprint available at reproducible.cs.princeton.edu.
[13] Joelle Pineau et al. 2020. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." arXiv preprint arXiv:2003.12206.

to 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. 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.[14] 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 the 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 Strategic Plan prioritizes the support of systematic reviews of published research across fields adopting machine learning methods to address the reproducibility crisis in ML-based science. The Strategic Plan could also incentivize work on the creation of computational reproducibility infrastructure and a reproducibility clearinghouse that sets up benchmark datasets for measuring progress in scientific research that uses AI and ML.[15] Finally, the Strategic Plan could make government funding conditional on disclosing research materials, such as the code and data, that would be necessary to replicate a study. A similar step is already underway for NIH funded studies.[16] Taken together, these steps would lead to significant strides towards the aim of promoting transparent, effective, and responsible research.

---

[14] Leonard P. Freedman , Iain M. Cockburn, Timothy S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." PLoS Biology 13(6).

[15] Benjamin Haibe-Kains et al. 2020. "Transparency and reproducibility in artificial intelligence." Nature 586, E14–E16.

[16] "Final NIH Policy for Data Management and Sharing". Notice Number: NOT-OD-21-013. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Medicine	Bouwmeester et al.	2012	71	27	No train-test split
Neuroimaging	Whelan et al.	2014	—	14	No train-test split; Feature selection on train and test set
Autism Diagnostics	Bone et al.	2015	—	3	Duplicates across train-test split; Sampling bias
Bioinformatics	Blagus et al.	2015	—	6	Pre-processing on train and test sets together
Nutrition research	Ivanescu et al.	2016	—	4	No train-test split
Software engineering	Tu et al.	2018	58	11	Temporal leakage
Toxicology	Alves et al.	2019	—	1	Duplicates across train-test split
Satelitte imaging	Nalepa et al.	2019	17	17	Non-independence between train and test sets
Clinical epidemiology	Christodoulou et al.	2019	71	48	Feature selection on train and test set
Brain-computer interfaces	Nakanishi et al.	2020	—	1	No train-test split
Histopathology	Oner et al.	2020	—	1	Non independence between train and test sets
Computer security	Arp et al.	2020	30	30	No train-test split; Pre-processing on train and test sets together; Illegitimate features; others
Neuropsychiatry	Poldrack et al.	2020	100	53	No train-test split; pre-processing on train and test sets together
Medicine	Vandewiele et al.	2021	24	21	Feature selection on train-test sets; Non-independence between train and test sets; Sampling bias
Radiology	Roberts et al.	2021	62	62	No train-test split; duplicates in train and test sets; sampling bias
IT Operations	Lyu et al.	2021	9	3	Temporal leakage
Medicine	Filho et al.	2021	—	1	Illegitimate features
Neuropsychiatry	Shim et al.	2021	—	1	Feature selection on training and test sets

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

4. Build and maintain infrastructure designed to independently test the validity of the claims of AI performance across applications under Strategy 6.

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 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 differ from their stated purpose. This is especially true when such applications dictate 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 speech patterns to determine candidates' personality traits or competencies from short video interviews and function as "algorithmic pre-employment assessments" to make hiring decisions easier. But it is generally understood by experts that these tools have significant shortcomings when it comes to predicting actual job performance. Nevertheless, Raghavan et al. describe how 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.[17]

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.[18] [19] However, this does not

[17] 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.
[18] 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).
[19] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." Science advances, 2018.

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 to address these important social issues.

As a result, we see evaluating validity as a core component of ethical and responsible AI research and development. The strategic plan could support such efforts by prioritizing funding for setting standards for and making tools available to independent researchers to validate claims of effectiveness of AI applications.

5. Incentivize and promote effective data stewardship under Strategy 5.

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. Such datasets raise serious privacy concerns and they may be used to support research that is counter to the intent of the people who have contributed to them. The Strategic Plan 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 Strategic Plan 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.[20] 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 the same vein as our point about downstream impacts (Section 2), existing ethical oversight mechanisms within academia such as IRBs 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

[20] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. "Mitigating dataset harms requires stewardship: Lessons from 1000 papers." NeurIPS 2021 (Datasets and Benchmarks track).

under the purview of IRBs. This compounds issues with runaway datasets and exacerbates ethical concerns with the creation and use of datasets.

The Strategic Plan can address this gap by supporting the development of 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. Finally, the Strategic Plan 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.[21]

* * *

We appreciate the opportunity to provide these comments and welcome the opportunity to discuss any questions.

Respectfully submitted,

Sayash Kapoor
Graduate Student, Department of Computer Science, Princeton University

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

Solon Barocas
Principal Researcher, Microsoft Research and Adjunct Assistant Professor, Department of Information Science, Cornell University

Arvind Narayanan
Associate Professor of Computer Science, Princeton University

Contact: 609-258-5306; mihir@princeton.edu

[21] 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.