National Science Foundation (NSF)
Docket No. 2024-19245
Request for Information on the CHIPS and Science Act Section 10343. Research Ethics

November 15, 2024

Dear NSF colleagues,

We, a group of scholars affiliated with Stanford's Ethics and Society Review (ESR) and the Stanford Institute for Human-Centered Artificial Intelligence (HAI), offer the following submission in response to your Request for Information on the CHIPS and Science Act Section 10343.

In 2020, the ESR was established by a multi-institute coalition within Stanford with the explicit goal of helping researchers consider and mitigate the ethical and societal risks posed by their research beyond those covered by an Institutional Review Board (IRB).[1, 2] Our response draws on our five years of experience operating the ESR ethical reflection process as a requirement for multiple grant cycles at HAI. To date, ESR has worked with more than 250 researchers to consider the ethical and societal implications of their research, which corresponds to over 160 proposals reviewed and equates to over $67 million in research funds for AI-related research.

We believe that all researchers and funding agencies have a responsibility to mitigate potential long-term harms from their research. We provide select examples, on-the-ground experiences, and perspectives from devising and administering an ethical reflection process for research. While we believe that ethical reflection can be integrated into any grantmaking process, the exact process can and should be tailored to the needs of the institution.

First, we draw lessons from our ethical reflection process to make the case that promoting ethical and societal reflection within the NSF's grantmaking is not only possible, but an ideal point in the research process to begin doing so.

Second, we describe our approach to incorporating ethical reflection into a grant call and explain how and why our process has changed since 2020, including administrative challenges that come with establishing this type of process and our current solutions.

---

[1] Bernstein, Michael S., Margaret Levi, David Magnus, Betsy A. Rajala, Debra Satz, and Quinn Waeiss. 2021. "Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding." Proceedings of the National Academy of Sciences 118(52): e2117261118. doi:10.1073/pnas.2117261118.
[2] Stanford HAI, "A New Approach to Mitigating AI's Negative Impact," June 24, 2021, https://hai.stanford.edu/news/new-approach-mitigating-ais-negative-impact.

Third, we highlight the common ethical issues that have arisen as part of our review of AI research. It is important to note that while the issues described in this response are an accurate depiction of issues currently present in AI research, it will not be relevant forever. As AI research progresses, so will the ethical issues that come from that research.

We address the specific questions posed in the request for information below.

Question 1: Describe ethical, social, safety, and/or security risks from current or emerging research activities that you believe might be of concern to the community, profession, or organization with which you are connected.

During the five years we have spent reviewing hundreds of research proposals as part of Stanford's ESR, we have observed several crucial risks that come with AI research activities. As part of our continuous efforts to improve the ESR process, every proposal and the accompanying panelists' comments are content coded for the proposal's substantive focus, the risks raised, and mitigation strategies put forward. From this, we have been able to derive quantitative measures of the most common salient ethical and societal risks associated with AI research. These are the top three risks that arise in the projects we review:

1. **Bias (80% of Stanford HAI research projects):** The quality of an AI tool stems in part from the quality of the underlying training data and algorithm guiding tool outputs. These data and outputs are often biased along racial, ethnic, gender, and socioeconomic lines. Bias, therefore, serves as a first-order concern in AI development. If the information informing the development of an AI tool is biased, additional risks and harms are likely to follow.

2. **Motivated misuse (61% of Stanford HAI research projects):** New technologies often give rise to new capabilities, and the proliferation of AI in particular offers tools attractive for motivated actors to misuse. Such risks are especially concerning in domains like generative AI and AI surveillance technologies, where nefarious actors could easily co-opt these tools for harm. Often such misuse risks are difficult to prevent once technology has made its way out of the lab.

3. **Exacerbating inequities (60% of Stanford HAI research projects):** Whether through biased datasets, algorithmic models that perform differently across subpopulations, or inappropriate implementation of AI tools, these new technologies risk further exacerbating societal inequities, especially when applied in domains where such inequities are already stark—e.g., healthcare, education.

The other most common risks that our expert panel identifies in projects include erosion of privacy as AI surveillance and data collection proliferates (45%); harms due to AI tool/model errors, including hardware and software malfunction (34%); excluding impacted communities from AI design, development, and implementation (26%); and user error due to misapplication or misinterpretation of the tool (23%).

Question 2: Which products, technologies, and/or other outcomes from research do you think could cause significant harm to the public in the foreseeable future?

Through reviewing hundreds of proposals, we have found that it is more productive to examine technologies within the specific contexts in which they operate rather than focusing solely on the technologies themselves. This approach recognizes that different contexts bring unique sets of norms that shape the evaluation of technological risks and harms. For instance, norms regarding information flow (i.e., privacy) differ across medical, educational, and professional settings. Hence, to analyze societal and ethical issues with technologies, we advocate for investigating how technologies interact with the specific contexts where they are applied and, in doing so, we shift our focus from the technologies per se to what happens when the research is implemented in the real world.

This shift in focus has illuminated features of technological research and development that, when intersecting with domains like healthcare, social services, education, and policymaking could lead to pernicious harms. These include AI decision support tools integrated into complex workflows, AI diagnostic tools, and generative AI applications.

For example, one project aimed to create an AI diagnostic tool for certain cancers. The researchers and panelists considered what could happen if such a tool were incorporated into a healthcare system. This raised concerns about insurance companies misusing such a diagnostic tool to deny insurance coverage claims. Researchers also contended with known disparities regarding cancer treatment and outcomes based on racial and socioeconomic demographics. Panelists were concerned that such disparities could be reified in model development, resulting in a biased diagnostic tool and healthcare decisions that could further exacerbate health inequities.

In another project, researchers aimed to develop a large language model (LLM) to support teachers' customization of curricular content to the needs of individual students. The researchers identified that their tool, if not properly tuned to the diverse needs of learners, could result in ineffective curricular content and further marginalization of struggling students. They also discussed ways to mitigate against concerns that an LLM tool could disadvantage teachers with limited technological experience.

Question 3: Describe one or more approaches for identifying ethical, social, safety, and/or security risks from research activities and balancing such risks against potential benefits.

Two approaches have been key for the ESR to help identify downstream societal risks of research and to balance such risks against potential benefits:

1. Empowering researchers to identify and mitigate risks in their own work

This takes the form of an initial ESR statement template that asks researchers to discuss the type of risk that could arise in their work, who could be affected, and *specific mitigation strategies* the researchers will implement in response to the risk. We have found that a well-designed ESR statement template is critical to providing researchers, who may not have ethical training, with common concerns and questions to consider as they develop their ESR statement (e.g., whose interests are represented in the project? Whose are excluded? What could go wrong if the model or tool malfunctions during deployment?). We developed these questions to help researchers better pinpoint the decisions on their project that could give rise to certain risks and capture how those risks could manifest in society.

Researchers are making risk/benefit judgments in their ESR statement development as they determine the necessary and appropriate strategies for mitigating risks they identified.

2. Collaboratively coaching researchers to further issue-spot and mitigate within their work

After researchers submit their ESR statement, our interdisciplinary panel coaches researchers on their risk identification and commitment to mitigation. Specifically, panelists provide feedback on: (1) the appropriateness of the Principal Investigators' proposed mitigation strategies; (2) additional project details necessary to assess the risks/benefits of the proposal; and (3) additional risks panelists identified in their review.

ESR panelists also ask the researchers targeted questions and outline why additional risks may remain on a project. This feedback is explicitly designed to provide researchers with additional scaffolding for issue-spotting and mitigation in the future. In cases where the reasonable, or even feasible, mitigation strategies are insufficient for the level of risk posed by the project, the ESR works with the project team to understand why we believe the perceived benefits do not outweigh the risks. We then inform the funding agency of that decision.

Say, for example, researchers propose the development of an AI decision support tool for clinicians and recognize that clinicians' uncritical and excessive reliance on the tool could harm patients. They suggest adopting the design principle of "augmenting, not replacing" clinician decision-making to protect against automation bias. ESR panelists might express concern that such a principle, on its own, would be insufficient to address the issue. They might ask for more

information on how the principle is explicitly instantiated in the tool design and implementation. Finally, ESR panelists might raise concerns about how patients with complex medical issues need more attention from clinicians and how overreliance on the AI tool could harm patients' diagnostic journey.

In another example, researchers proposed developing a toxicity prediction algorithm for new chemicals. Using the example risk and mitigation strategies provided in the ESR statement prompt, the researchers found that their tool could be misused to identify and develop toxic chemicals. To mitigate this risk, they committed to publicizing the appropriate and intended uses of their tool to users and implementing oversight mechanisms to monitor the use. The ESR panelists encouraged the researchers to develop more extensive mitigation measures due to the tool's attractiveness to malicious users. The researchers consulted with federal policymakers, industry leaders, and other scientists, and concluded that existing guidance remained inadequate.[3] They voluntarily sought further guidance from the ESR, sparking the ESR's development of a mitigation framework for addressing misuse of AI in biomedicine.

Question 4: Describe one or more strategies for encouraging research teams to incorporate ethical, social, safety, and/or security considerations into the design of their research approach. Also, how might the strategy vary depending on research type (for example, basic vs. applied) or setting (for example, academia or industry)?

By requiring that researchers complete the ESR process prior to the release of grant funds, the ESR addresses the self-selection bias—where only motivated individuals incorporate ethical considerations—and ensures that all grant applicants engage with ethical reviews. In addition to the standard materials submitted for funding, applicants must provide an ESR statement. This statement outlines potential societal and ethical issues associated with their projects, as well as preliminary strategies for addressing them. This requirement ensures that applicants are not only identifying potential risks but are also committed to implementing proactive mitigations that reduce downstream harms.

Then, throughout the ESR process, our panelists and applicants engage in an iterative and cooperative approach, providing project teams with the necessary cognitive scaffolding to further specify ethical risks and fine-tune mitigation strategies.

As they receive feedback from the ESR panel, the team iterates on the research design, their mitigation commitments, and the set of stakeholders they engage.

[3] Shankar, Sadasivan, and Richard N. Zare. 2022. "The Perils of Machine Learning in Designing New Chemicals and Materials." Nature Machine Intelligence 4(4): 314–15. doi:10.1038/s42256-022-00481-9.

The ESR evaluation process fosters an ethical mindset among research teams during the design phase of their projects, influencing not only their current work but also future endeavors. By requiring ESR statements, teams are prompted to consider ethical implications early on, potentially leaving a lasting impact on their approach to research. The iterative feedback process with our panelists—comprising experts from various disciplines—serves as a form of coaching, teaching teams how to identify overlooked risks and assess the adequacy of their mitigation strategies.

Question 6: How could ethical, social, safety, and/or security considerations be incorporated into the instructions for proposers or into NSF's merit review process? Also, what challenges could arise if the merit review process is modified to include such considerations?

The NSF is uniquely positioned to incorporate ethical and societal considerations into instructions for proposers. Researchers must already include in their NSF proposal package a Broader Impacts statement, where they reflect on the potential of their work to benefit society. It would be a logical extension of this section of the proposal to ask researchers to reflect on the potential of their work to cause harm in society. The NSF Proposal & Award Policies & Procedures Guide (PAPPG) instructions could expand on the list of societal benefit examples to identify their inverse (e.g., harming individuals' well-being, disenfranchising marginalized groups from STEM participation, decrease trust in science).

Soliciting this reflection at the point of proposal both socializes researchers around these issues as part of NSF review criteria and prioritizes these issues at the development of the research rather than as an afterthought. Furthermore, including a societal harm component could bolster expectations for evaluating success: A project's success is judged not only by its benefit to society but also whether it has contributed to any societal harms.

Similarly, the NSF could incorporate such considerations directly into the merit review process. Procedurally, this could require additional reviewer(s) to provide input explicitly on the broader impacts—especially potential harms—of the proposed work. Program officers may wish to recruit reviewers with expertise at the intersection of societal structures and the substantive focus of the project to ensure holistic evaluation of the project's intended and potential impacts. Additional criteria, mirroring review criteria that have already been established by the NSF, could help focus reviewers' evaluation of ethical and societal risks. Such criteria should include:

- What is the potential for the proposed activity to harm society or advance harmful societal outcomes?
- Does the project plan incorporate a mechanism to assess success and evaluate potential harms?

- How well-qualified is the individual, team, or institution to assess and mitigate the societal impact of their research?

By incorporating these additional criteria into the merit review process, the NSF can ensure balanced consideration of both the potential benefits and potential harms of proposed projects.

Some of the challenges our team encountered when implementing the ESR process can help inform how the NSF proceeds with incorporating ethical, societal, and safety considerations into their review process. Adding new review requirements increases the time and expertise needed for review, but the NSF can leverage an existing strategy to address this: recruiting ad hoc reviewers for projects with tailored ethical assessment needs. Updated review criteria can also help scope assigned reviewers' assessments. Given the already resource-intensive nature of NSF merit review panels, it may take less time to integrate one or two additional panelists into an existing merit review rather than stand up an entirely separate process.

At the same time, adding societal harm considerations to the merit review process can also inform researchers' development of robust evaluation procedures to accurately capture the societal impacts of their work. This can further ensure the NSF is responsibly funding projects with the potential to improve society and without their projected benefits being offset by unmitigated harms.

Sincerely,

Quinn Waeiss
Research Affiliate, Ethics and Society Review, McCoy Family Center for Ethics in Society
Postdoctoral Fellow, Stanford Center for Biomedical Ethics
Stanford University

Raio Huang
Tech Ethics and Policy Rising Scholars Research Associate, Ethics and Society Review, McCoy Family Center for Ethics in Society
Stanford University

Betsy Arlene Rajala
Program Director, Ethics and Society Review, McCoy Family Center for Ethics in Society
Stanford University

Michael Bernstein
Faculty Co-Chair, Ethics and Society Review, McCoy Family Center for Ethics in Society
Associate Professor of Computer Science
Faculty Affiliate, Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Margaret Levi
Faculty Co-Chair, Ethics and Society Review, McCoy Family Center for Ethics in Society
Professor Emerita of Political Science
Senior Fellow, Center on Democracy, Development and the Rule of Law, Freeman Spogli Institute
Stanford University

David Magnus
Faculty Co-Chair, Ethics and Society Review, McCoy Family Center for Ethics in Society
Thomas A. Raffin Professor of Medicine and Biomedical Ethics, Professor of Pediatrics
Associate Dean of Research
Stanford University

Debra Satz
Faculty Co-Chair, Ethics and Society Review, McCoy Family Center for Ethics in Society
Vernon R. and Lysbeth Warren Anderson Dean of the School of Humanities and Sciences
Marta Sutton Weeks Professor of Ethics in Society, Professor of Philosophy
Stanford University