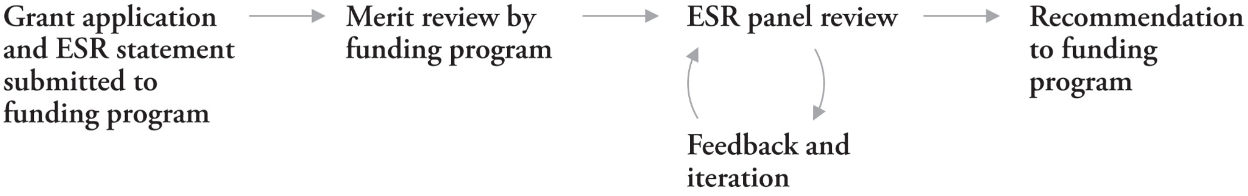
Final Report for NVF-PITU-Stanford University-Subgrant-012849-2020-11-18

In our Public Interest Technology University Network Challenge proposal, we proposed the creation of the Ethics and Society Review (formerly known as an Ethics Review Board), a process wherein grant proposals at a major AI institute undergo review from colleagues versed in technology and ethics, helping guide the researchers in articulating potential societal impacts and mitigating negative outcomes.

The ESR is an institutional process that guides technologists, who are often outside the purview of IRBs, in thinking about the long-term societal impacts of their work. The initial objectives of the ESR were to: (1) design a process that involves ethicists, social scientists and technologists in providing productive feedback to research projects at their proposal stage; (2) integrate that process as a requirement to receive funds from a major on-campus grant-giving institute; (3) gather evidence from grant PIs, ESR panel members, and institute directors as to the effectiveness of the process, including the actual eventual effects of the approved projects.

Since receiving support from the PIT-UN Challenge, the ESR has achieved its first objective by designing an iterative feedback process that incorporates a panel review by relevant experts in ethics from disciplines including anthropology, biomedical data science, biomedical ethics, communication, computer science, history, management science and engineering, philosophy, political science, and sociology. The graphic below illustrates the timing of this review process: after the successful merit review by the funding program, proposals and their ESR statements are reviewed by the ESR panel for potential ethical and societal consequences of the research. The panel provides feedback to the researchers and, in some cases, requests that the researchers iterate their ESR statement to address issues raised by the ESR panel. When the process is complete, the ESR submits its recommendation to the funding program, and funds are released to the researchers.


Grant application and ESR statement submitted to funding program
Merit review by funding program
ESR panel review
Recommendation to funding program
Feedback and iteration

Regarding the second objective, the ESR has successfully integrated into the grant-making process at a major on-campus grant-giving institute (Human-Centered Artificial Intelligence) for the past two years. We are also in the process of integrating the ESR into other grant-giving institutes at Stanford University that focus on sustainability and social sciences.

Relating to the third objective, in our pilot evaluation of the 2020 ESR, surveys and interviews of researchers who interacted with the ESR found 100% (95% CI: 87 to 100%) were willing to continue submitting future projects to the ESR, and 58% (95% CI: 37 to 77%) felt that it had influenced the design of their research project. Nearly all interviewees reported that the ESR process encouraged them to think more deeply about the broader implications of their research. Eight interviewees said that the ESR process raised new issues for them to think about. For six others, while the process did not raise new issues, it encouraged them to deepen their reflection on ethical implications that they were already considering. Many reported that the forcing function of the ESR statement and the panel's feedback led participants to discuss the issues with others, which revealed new issues.

In our evaluation, we also identified areas for improvement in future iterations of the ESR. In particular, while the ESR statement prompt was kept brief, participants requested more specificity, including additional examples of ethical violations in research, with some even proposing a workshop to help clarify the rubric to be used in evaluating the seriousness of a risk. Separately, ESR panelists and directors raised concerns about the level of commitment required by faculty panelists and suggested a first-round triage from doctorate-level staff with expertise in ethics, with the goal of identifying grants that do not require escalation for faculty panel review.

To strike a path forward, the ESR drew on data from its first year. Instructions for new iterations of the process now include a list of the most common risks raised by panelists, as well as example principles for mitigation and resulting research designs, for each risk. These examples serve as benchmarks to set expectations for the researchers as well as the panelists. The ESR also implemented a first-round triage with four doctorate-level staff, reducing the number of proposals that required faculty review by 50%.

The first-year evaluation of the ESR is described in the following publication:

- Bernstein, Michael, Margaret Levi, David Magnus, Debra Satz, Betsy A. Rajala, and Charla Waeiss. "Ethics and society review: Ethics reflection as a precondition to research funding," *Proceedings of the National Academy of Sciences* 118(52): 1-8.

The ESR has been discussed in the following media publications:

- https://hai.stanford.edu/news/new-approach-mitigating-ais-negative-impact
- https://humsci.stanford.edu/feature/qa-new-ethics-and-society-review-addresses-ethical-and-societal-impacts-proposed-research

All Stanford University activities conducted with the Grant funds were and are consistent with charitable purposes as set forth in Section 501(c)(3) of the Internal Revenue Code, and Stanford University complied with all provisions and restrictions contained in this Agreement, including, for example and without limitation, those provisions relating to lobbying and political activity.