Ethics and society review: Ethics reflection as a precondition to research funding

Michael S. Bernstein[a], Margaret Levi[b,c,1], David Magnus[d], Betsy A. Rajala[b], Debra Satz[e], and Quinn Waeiss[b]

[a]Department of Computer Science, Stanford University, Stanford, CA 94305; [b]Center for Advanced Study in the Behavioral Sciences, Stanford University, Stanford, CA, 94305; [c]Department of Political Science, Stanford University, Stanford, CA 94305; [d]Department of Pediatrics, Stanford University, Stanford, CA 94305; and [e]Department of Philosophy, Stanford University, Stanford, CA 94305

Contributed by Margaret Levi; received September 20, 2021; accepted November 4, 2021; reviewed by David Danks and Rose McDermott

Researchers in areas as diverse as computer science and political science must increasingly navigate the possible risks of their research to society. However, the history of medical experiments on vulnerable individuals influenced many research ethics reviews to focus exclusively on risks to human subjects rather than risks to human society. We describe an Ethics and Society Review board (ESR), which fills this moral gap by facilitating ethical and societal reflection as a requirement to access grant funding: Researchers cannot receive grant funding from participating programs until the researchers complete the ESR process for their proposal. Researchers author an initial statement describing their proposed research's risks to society, subgroups within society, and globally and commit to mitigation strategies for these risks. An interdisciplinary faculty panel iterates with the researchers to refine these risks and mitigation strategies. We describe a mixed-method evaluation of the ESR over 1 y, in partnership with an artificial intelligence grant program run by Stanford HAI. 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. The ESR panel most commonly identified issues of harms to minority groups, inclusion of diverse stakeholders in the research plan, dual use, and representation in datasets. These principles, paired with possible mitigation strategies, offer scaffolding for future research designs.

ethics | machine learning | computer science | societal consequences

Whether research diffuses into society through technological adoption, through field experiments, or through policy, researchers must reflect on how to identify and mitigate the risks that the diffusion of their work presents to human society. These risks include, for example, the possibility that their contributions to artificial intelligence (AI) might exacerbate biases in the criminal justice system (1), that their urban planning concepts might backfire when implemented (2), or that their elections research might influence electoral outcomes (3). Through these projects and many others, researchers must grapple with not just the benefits of their work but also the risks that their work presents to society: to forms of social organization ranging from groups to nations to humanity as a whole.

Research ethics review often focuses on risks to human subjects, not risks to human society, placing societal risks out of scope and out of jurisdiction. In the United States, ethics review is associated with Institutional Review Boards (IRBs) and is governed by the Common Rule (4, 5). The Common Rule gives IRBs jurisdiction over risks to human subjects,* who are the individuals directly engaged or studied in the research. However, the Common Rule governing IRBs specifically disallows review of consequences to human society: "The IRB should not consider possible long-range effects of applying knowledge gained in the research [. . .] as among those research risks that fall within the purview of its responsibility" (5). This regulation is generally interpreted to mean that IRBs should decline to review research risks to human society.

It is not unreasonable to worry about IRB overreach—almost every action carries potential risks of harms—yet it is inappropriate to ignore the risks that research poses for our collective future: the risks of AI to the future of work, the risks of sustainability interventions to the societies that they are purported to support, the risks of the internet to professional media and accurate information. In the light of these risks, recent scholarship has argued that research carrying substantial societal risk should undergo ethics review. One thread of this scholarship proposes to expand the definition of "human subject" to include societies (6), and IRBs such as the Microsoft Research Ethics Review Program have adopted this expanded purview (7). An alternative approach directly calls for expanding the Common Rule to include Respect for Societies as a principle (3), or revising it to address substantive ethical issues rather than procedural concerns (8). A third approach directly seeks to regulate some fields to require ethics reviews or audits (9–11) or enforces it during peer review (12, 13). Another integrates ethics training into laboratory meetings (13) and course curricula (14–16). A final approach focuses on articulating ethics guidelines by researchers (17–19) or by professional associations (20–22). The goal of our

Significance

Research fields that hold transformative possibilities for improving the human condition also raise risks of negative ethical and societal outcomes. These ethical and societal risks fall outside the purview of most research reviews. We introduce an iterative review process that draws these fields into reflection and mitigation of ethical and societal risks by conditioning access to grant funding on completion of the process. A 1-y evaluation of our approach with an artificial intelligence funding program at our university suggests that this approach is well-received by researchers and positively influenced the design of their research. This process has also generated lists of common risks and mitigation strategies, to provide scaffolding for future processes.

Author contributions: M.S.B., M.L., D.M., B.A.R., D.S., and Q.W. designed research; M.S.B., M.L., B.A.R., and Q.W. performed research; M.S.B., B.A.R., and Q.W. analyzed data; and M.S.B., M.L., B.A.R., and Q.W. wrote the paper.

Reviewers: D.D., University of California San Diego; and R.M., Brown University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: mlevi@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117261118/-/DCSupplemental.

Published December 21, 2021.

*Our focus in this article is on research that may impact human subjects or human societies. IRBs' purviews also consider other issues, such as animal experimentation or biospecimens, for other areas of research.

work is to leverage these conceptual and organizational insights to design a concrete process that engages researchers whose work typically falls outside the purview of their institution's current review processes.

We introduce Ethics and Society Review (ESR), a process that facilitates ethical and societal reflection as a requirement to access funding. With the ESR, grant funding from participating institutions is not released until the researchers successfully complete an iterative review process on their proposed project. Conditioning funding on the ESR process helps engage researchers at the formative stages of their research, when projects are still open to change, and ensures broad engagement with the process rather than self-selection of just those who are motivated.

For funding organizations that incorporate the ESR in their grant process, researchers submit a brief ESR statement alongside their grant proposals. The ESR statement describes their project's most salient risks to society, to subgroups in society, and to other societies around the world (see *Materials and Methods*). This statement articulates the principles the researchers will use to mitigate those risks and describes how those principles are instantiated in the research design.

After the funding program conducts its grant merit review, it sends only the grants recommended for funding to the ESR for ethics review (Fig. 1). The ESR convenes an interdisciplinary panel of faculty that considers the studies' risks and mitigations in the context of possible benefits to society and determines the adequacy of the ESR statement provided by the investigators. Its goal is not to eradicate all potential or actual negative impacts—which is often impossible—but to work with the researchers to identify negative impacts and to devise reasonable mitigation strategies. Over 1 to 2 wk, the faculty panel engages in iterative feedback to the researchers, which can include raising new possible risks, helping identify collaborators or stakeholders, and brainstorming additional mitigation strategies. Principal investigators (PIs) submit written responses to the ESR feedback as addenda to their original statement. These addenda can include replies to the panel's feedback as well as commitments to specific mitigation strategies.

When the process is complete, the ESR submits its recommendation to the funding program, and funds are released to the researchers. *Materials and Methods* describes this process in additional detail, and *SI Appendix* includes the prompts used. For a comparison of the IRB process with the ESR, see Table 1.

We initiated the ESR in the context of AI research, in partnership with a grant program run by the Stanford Institute for Human-Centered Artificial Intelligence (HAI). This context serves as a useful test case for the ESR for several reasons. First, AI research is often outside the scope of IRB review, yet AI is wrestling with the ethical and societal implications of its work. AI systems are implicated in generating and propagating disinformation (23–25), depressing wages for workers (26–29), perpetuating systemic inequality in policing and the justice system (1, 30, 31), and advancing unequal healthcare outcomes (32). Among the challenges are oversights in who is and is not represented in the dataset (33), who has a seat at the table in the design and deployment of the AI (34), who is intended to benefit and be harmed by the AI (35), and what likely consequences might arise (36, 37). AI systems have become embedded into sociotechnical systems where their direct and indirect impacts now reinforce racism, entrench economic disparities, and facilitate other societal ills (1, 35, 38–41). The AI grant program we partnered with attracts researchers from many areas, including the arts, Earth science, humanities, medicine and social science—not only engineering. While some of the ethical issues raised in the ESR process are particularly salient in AI research (e.g., publicness), nearly all of the issues raised apply to a wide range of disciplines (e.g., representativeness, diverse deployment, and design).

We use brief anonymized cases from the ESR deployment to illustrate the ESR process; more detailed case studies are in *SI Appendix*. We discuss a project by faculty in Medicine and Electrical Engineering focused on noninvasive stress sensing at work. The ESR statement expressed risks about employers using this technology to surveil and depress the status of workers. In response, the ESR asked for principles to mitigate this risk and specific design decisions that the researchers would be making in line with those principles. The researchers committed to building a privacy-preserving architecture for the tool and emphasizing this architecture and its importance in writing and presentations on the research. We also discuss a project by faculty in Earth Science and Computer Science who proposed remote-sensing models for sustainability applications. In their ESR statement, they identified risks including that the models might perform differently in different parts of the world, and they committed to auditing their models globally, specifically focusing development on Africa, to challenge the status quo of similar models focusing on the United States. (The ESR panel did not request iteration, given these commitments.) In the third case we discuss, when faculty from Education, Psychology, and Computer Science proposed a reinforcement learning AI system to support student retention, the ESR pointed out that the AI might minimize its loss function by focusing on the learners who it is most likely to be able to retain rather than those most at risk. The researchers responded by highlighting a coinvestigator who studies inclusive educational experiences for marginalized groups and committing to evaluate the system to test for this risk.

We report on a year-long mixed-method evaluation of the ESR at Stanford University, during which time it reviewed 41 grant proposals. We surveyed and interviewed lead researchers on these projects to understand their experiences with the ESR and conducted an inductive analysis of the ESR statements and panel feedback.

Results

In collaboration with the Stanford Institute for Human-Centered Artificial Intelligence (HAI), the ESR reviewed 6 large grants ($2.5 million over 3 y) and 35 seed grants ($75,000 over 1 y). ESR panelists asked the researchers from all 6 of the large grants (100%) and 10 of the seed grants (29%) to iterate based on the ESR's feedback, of which 3 of the seed grants (9%) iterated multiple times. All were eventually supported by the ESR, not as risk-free but as having appropriate mitigation plans.

We surveyed the lead researchers from the 35 seed grants that engaged with the ESR's process. The survey investigated researchers' prior exposure to ethics reflection and review, the level of influence that the ESR feedback had on the project, the aspects of the process that the researchers found most helpful and least helpful, and opinions on whether the ESR can help mitigate negative outcomes. We also conducted semistructured

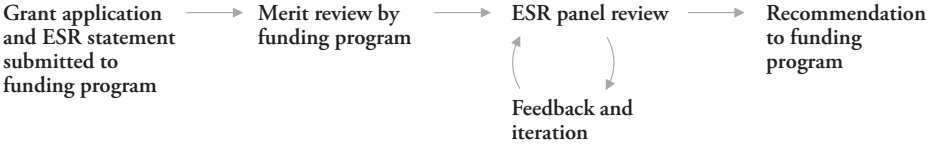

Fig. 1. The ESR process accepts initial statements from researchers when they submit the grant then iterates with them prior to releasing funding.

Table 1. The IRB is focused on risks to human subjects, whereas ESR is focused on risks to society, groups within society, and to the world

	IRB	ESR
Focus	Significant risks to human subjects	Significant risks to societies, to groups within those societies, and to the world
Requirement	Data collection cannot begin, and funds cannot be spent to pay research participants, until IRB protocol is approved	Grant funding cannot be released by the funding program until the ESR process has completed
Submission	Specifics of research design, including any procedure followed by research participants and any materials shown to research participants	Goal of research, direct and indirect stakeholders, and higher-level research design. Articulation of principles to mitigate negative outcomes and description of how those principles are instantiated in the research design
Timing	Regular (e.g., monthly) deadline	Synchronized with grant funding programs
Possible outcomes	Approval, return with comments, (rare:) rejection	Approval, return with comments, request synchronous conversation, (rare:) rejection
Amendment and renewal	Protocols must be amended if the research design changes, and expire after a fixed period (e.g., 3 y)	Protocols will be examined annually as part of the researcher's annual grant report to the funding institution

To enable engagement with the ESR early in the research lifecycle, researchers work with the ESR prior to funding's being released.

follow-up interviews with lead researchers. The survey and interviews were both covered by an IRB-approved consent process, and the survey, interview instruments, and descriptive participant statistics are included in *SI Appendix*. All analyses were exploratory, so hypotheses were not preregistered and *P* values are not reported.

Overall, researchers wished to continue the ESR process. The survey asked participants whether they would submit to the ESR again. All were willing (Fig. 2; 95% CI: 87 to 100%). Stratifying the responses by whether grants were asked to iterate with the ESR, among those who did not iterate with the ESR 37% said they would only do it if required and the rest (63%) said they would do it voluntarily; among those who iterated with the ESR, all said they would do it voluntarily. Based on interviews, researchers generally expressed satisfaction with the ethical reflection process required by the ESR. Those who iterated particularly appreciated the engagement with panelists and the opportunity to commit to some detailed mitigation strategies for ethical concerns that arose in the process.

Fifty-eight percent (95% CI: 37 to 77%) of the self-reported responses indicated that the ESR process had influenced the design of their research project (Fig. 3). In the interviews, six researchers expressed that, rather than influencing specific components of their project, the ESR process shaped its entire development. One researcher referred to this as "ethics by design." Most projects did not iterate with the ESR, so the parts of the process they experienced were the writing of the ESR statement and reading the ESR panel's feedback. Among those who iterated with the ESR, 67% indicated that the ESR process had influenced their design.

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.

Overall, the ESR process also appeared to raise the consciousness of some researchers to engage more seriously with research ethics going forward:

> The [ESR statement] requirement . . . led me to engage with my co-PI . . . because, as a psychologist, I . . . wasn't aware of some of the potential ethical implications that this . . . AI work may have, and it helped me to engage with my co-PI as part of this requirement. — Researcher, social science

> In fact, we might flip our whole research approach to being about privacy. . . . [The] pretty strong reaction from the [ESR made] us rethink, to lead with . . . privacy. We really just want buildings to be spaces that people flourish in and we need to do it in some way that's going to be the most privacy-preserving [as] possible . . . We don't have answers yet, but . . . it's definitely helped us think about a better way to approach the research, how we're doing it and how we're talking about it. — Researcher, engineering

When iterating with the ESR, researchers submitted addenda to their original ESR statement that addressed the feedback provided by panelists. This often included a commitment to additional mitigation strategies that were not outlined in the initial statement. The most common resulting change, encompassing 3 of the 10 proposals that iterated with the ESR, was a commitment to specific strategies for sharing their findings and promoting techniques that could prevent malicious or erroneous applications of their work. Other changes that researchers made to their iterated proposals include commitments to additional experiments before drawing conclusions about target populations; contextualizing feedback provided by an AI tool to maintain motivation in students that could be harmed without it; holding training and sensitivity sessions with practitioners represented in medical data; auditing algorithmic performance and assessing for the need for additional samples; broadening research questions

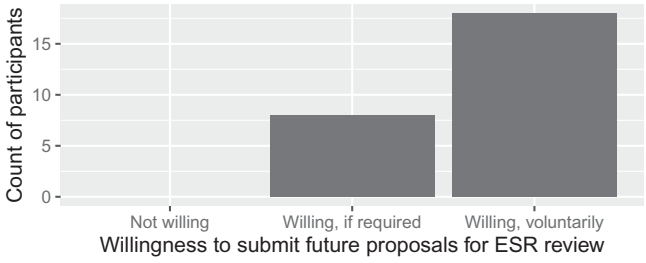

Fig. 2. All participants were willing to engage in the ESR process again.

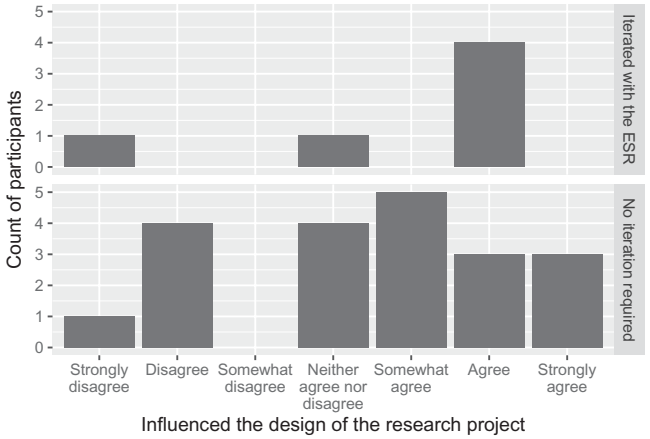

Fig. 3. Sixty-seven percent of researchers who iterated with the ESR, and 58% of all researchers, felt that the ESR process had influenced the design of their project.

to examine people's trust in AI-generated content; and employing and advocating for the use of data trusts.

The iterative process allowed PIs and panelists to engage in an ongoing conversation about the risks and appropriate mitigation strategies within the proposed research. For example, on one project PIs named risks related to representativeness in their initial statement, indicating that they will measure demographic representation in their training data, use diverse datasets wherever possible, and monitor the performance of their algorithm as it relates to the demographic groups in their data. Panelists raised an additional risk in response: diverse design and deployment. They recommended that the PIs get input from relevant policy/ethics experts for their algorithmic assessments and consider how to engage relevant stakeholders in the development of their tool. The panelists also requested additional mitigation strategies from the PIs to address representativeness concerns. They asked the PIs to elaborate on how they will address cross-cultural differences that are relevant for their data-labeling tasks and how they could detect cultural bias in their data. The panelists also recommended that the PIs track the diversity of their data annotators. The PIs clarified in their response that they were not seeking to define how cross-cultural interactions should be labeled; instead, they were striving to develop a tool that enables psychologists and behavioral scientists to address such questions. The PIs committed to tracking the demographic information of annotators where possible, highlighting where such tracking is not feasible (e.g., for existing third-party datasets), and the attendant limitations that follow from the demographic composition of their annotators or lack of such information. Following this iterative process, panelists and researchers were both ready for the project to move ahead, and the ESR recommended the proposal for funding.

Few researchers in our study had engaged in formal ethics review prior to the ESR. Nearly 80% of survey respondents self-reported that they had engaged in informal conversations about ethics within the month prior to the ESR process, and a majority of interviewees (10) mentioned engaging with research ethics frequently. However, only 8% of survey respondents had engaged in a structured ethics review beyond the IRB, and most interviewees reported their ethical reflections to focus on risks to individual human subjects and not broader risks to society.

Ultimately, researchers felt that the ESR process made it less likely that their project would misstep and wind up in the public eye for the wrong reason. Seventy-three percent of survey respondents agreed that the ESR reduced the probability of public criticism of their project, with 100% of those who iterated with the ESR agreeing. In the interviews, researchers indicated that, while they did not expect the ESR to shield them from warranted public criticism, the process had better prepared them for potential issues and appropriate ways to address them.

Issues Raised in ESR Feedback. One of the authors conducted open coding across all of the ESR statements and panelist responses using a grounded theory method to develop a set of 14 codes of themes brought up. These codes and their definitions are included in Table 2. A second author independently coded a subset of statements and panelist responses to test replicability; interrater reliability via Cohen's kappa averaged 0.96 per theme, with a range of 0.83 to 1. See *SI Appendix* for further details on the coding process and additional information on panelists' feedback to researchers.

The themes raised most frequently by the ESR panel (Table 2) were, in order of frequency, harms to subgroups (11), followed by diverse design and deployment (8), dual use (8), representativeness (6), and issues that fell under IRB purview (6). The issues raised most frequently by PIs were similar. The evaluation identified areas of improvement not only in the ESR process going forward but also in the IRB process: Both researchers and panelists raised issues that the IRB should cover, including how data are protected.

In 26 of the 35 seed grant projects, the ESR raised new themes that the PIs had not discussed in their ESR statements. In addition to raising new risks and continuing the conversation, some panelists also provided specific mitigation strategies. In some cases, a panelist raised a new issue and outlined possible mitigation strategies for it; in others, the researchers had raised the issue but left it insufficiently addressed. It was rarer for panelists to identify a potential collaborator for researchers to work with or refer the researchers to specific work on an issue.

Desire for Additional Scaffolding. Researchers wanted the ESR not only to push them to broaden their ethical and societal lenses but also to provide them with the scaffolding needed to navigate complex ethical and societal issues. 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.

> [The ESR didn't] really help us figure out how to address these [ethical issues]... . [They should] tell us how big the issues really are. . .the hard stuff is figuring out how important a particular ethical concern is. As researchers, we're often left with trying to decide whether the positives outweigh the negatives in terms of use cases and ethics. What I found that the [ESR] didn't do was really help us in making those decisions about whether the positives outweigh the negatives or not. — Researcher, medicine

> It'd be nice if there [were] some foundational or bedrock things that were in [the statement prompt]. You know, one risk is [the statement] becomes template-y, which I think is a risk and a problem. But having to write another page when you're an academic is useful because it forces you to think these things through, which we've discussed, but it's just more burden. In my view the burden here is worth it but [if] there [were] some sort of help that would scaffold a researcher through rather than just, "okay, here's a blank page, start from scratch." — Researcher, social science

Vesting Rejection Power in the ESR. We surveyed whether researchers felt that the ESR should be empowered to deny funding to a project. We expected this issue to be quite contentious, but

Table 2. Risk themes raised in the ESR process

Theme	Researcher statement frequency (n = 35 proposals)	Panelist response frequency (n = 35 proposals)	Refers to issues that pertain to. . .
Representativeness	18	6	Any risks or concerns that arise from insufficient or unequal representation of data, participants, or intended user population (e.g., excluding international or low-income students in a study of student well-being)
IRB purview	14	6	Any risks or concerns regarding the research that fall under IRB purview (e.g., participant consent, data security, etc.)
Diverse design and deployment	13	8	Incorporating relevant stakeholders and diverse perspectives in the project design and deployment processes (e.g., consulting with parents who have been historically disadvantaged to develop fairer school choice mechanisms)
Dual use	10	8	Any risks or concerns that arise due to the technology being coopted for nefarious purposes or by motivated actors (e.g., an authoritarian government employed mass surveillance methods)
Harms to society	10	5	Potential harms to any population that could arise following from the research (e.g., job loss due to automation)
Harms to subgroups	7	11	Potential harms to specific subgroup populations that could arise following from the research (e.g., technical barriers to using an AI that is prohibitive to poorer populations)
Privacy	4	1	Any risks or concerns related to general expectations of privacy or control over personally identifiable information (e.g., consequences of mass surveillance systems for individuals' control over their information)
Research transparency	3	0	Sufficiently and accessibly providing information such that others can understand and effectively employ the research, where appropriate (e.g., training modules for interpreting an AI model)
Accountability	2	2	Questions of assigning responsibility or holding actors accountable for potential harms that may arise (e.g., how to assign responsibility for a mistake when AI is involved)
Other	2	3	Other issues not covered above (e.g., intellectual property concerns)
Tool or user error	2	4	Any risks or concerns that arise from tool/model malfunction or user error (e.g., human misinterpretation of an AI model in decision-making)
Collaborator	1	1	Any risks or concerns that specifically relate to a collaborator on the research project (e.g., whether a collaborator could credibly commit to a project on inclusivity when their platform was notorious for exclusive and harmful behavior)
Methods and merit	1	2	Any risks or concerns reserved for methods and merit reviews of the grant proposal (e.g., whether model specifications are appropriate for the goals of the research)
Publicness	0	2	Questions of using publicly available data for research when those that generated the data are unaware of researchers' intended use of their data (e.g., use of Twitter data without obtaining express consent from affected Twitter users)

The researchers, in their ESR statements, were most likely to raise issues of representativness. The panelists, in their feedback, were most likely to raise issues regarding harms to subgroups. Both researchers and panelists also commonly focused on diverse design and deployment, dual-use concerns, harms to society, and issues pertaining to IRB purview.

there was generally a consensus via the survey that this was desirable, with no moderate or strong disagreement (Fig. 4). Among interviewees, although 11 agreed to varying degrees that the ESR should be empowered to reject an especially ethically problematic proposal, 5 of those participants strongly encouraged the ESR to prioritize the iterative process over a brute one-sided enforcement mechanism. Many believed that if a researcher does not demonstrate a willingness to engage with panelists' recommendations and feedback rejection of the project might be warranted, but only after a deliberative process of exchange between the ESR panel and the researchers.

Discussion

Evaluation of the first year of the ESR with a large, interdisciplinary AI program at our university suggests that the process can productively involve researchers in ethical and societal reflection early on in their projects. This is preferable, in our view, to dealing with these issues after the project has launched or is submitted for publication. In this section, we reflect on lessons from the evaluation, resulting changes to the ESR process, generalization of the ESR, and limitations of our study.

The evaluation feedback highlighted the tension that the ESR must navigate between providing structured criteria (e.g., checklists) and more case-specific feedback. In desiring more scaffolding, many researchers wished for more structure to the review process: lists of risks, levels of concern attached to each risk, and a process for knowing when a risk was mitigated sufficiently. On the other hand, researchers appreciated that the ESR was responsive to the particularities of each project. To strike a path forward, the ESR will draw on data from its first year. Instructions

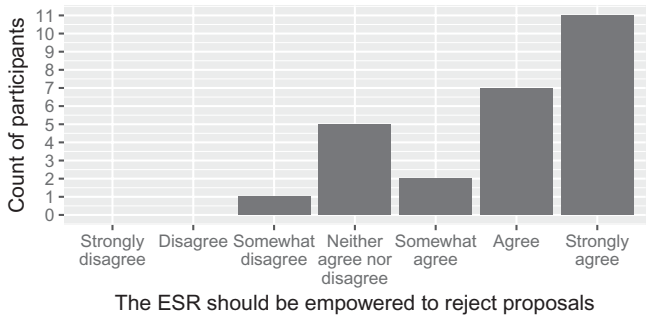

Fig. 4. Researchers were generally in favor of the ESR's being empowered to reject proposals if necessary.

for new iterations of the process, available in *SI Appendix*, 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.

Interest at the university level in expanding the ESR also raises the question of how to scale it from 40 grants per year to potentially over 100 grants per year, and how to ensure that it is sustainable. Articulating guidelines to aid researchers in the previous goal provides benefit here as well: new iterations of the ESR include a first round of triage from doctorate-level staff who have expertise in ethics, with the goal of identifying grants that do not require escalation for faculty panel review. Requiring some financial support from each partner program, staff triage can help the program from becoming too burdensome on a small set of faculty. Master's-level staff could also provide support during an initial triage round, especially if they undergo a calibration on the review process. For a description of the staff panel calibration and review process used in the second iteration of the ESR see *SI Appendix*.

Our model is not one where faculty who are trained in ethics point fingers at faculty trained in other areas. When possible, we pursue the metaphor of "coach" rather than "reviewer." This interactive model is a feature and not a bug of the process, as the ESR must navigate cultural change in the practice of research and translation across research fields.

How should the ESR handle conflicts of interest? These issues did not arise in our deployment, but neither did the ESR collect any information on, for example, whether outside funding or market opportunities might bias a researcher to focus on particular topics or populations. One approach would be to ask researchers to self-disclose any current or potential conflicts.

Can and how should the ESR transition from an *ex ante* review process to ongoing feedback? One intriguing possibility is that many funders require annual reports on their grants. We are currently coordinating with the funding program to request a brief update on the project in regard to the ESR procedures on which the panel and researchers agreed. Has the project changed in ways that would benefit from additional conversation or review? Are there unforeseen consequences that merit reconsideration? What happens when projects scale up or involve other partners, including those from the for-profit world?

We are currently expanding the ESR from AI research to other research areas and other funding programs. These programs include sustainability, bioengineering, and behavioral science projects involving community partners. Separate faculty panels must be recruited for these programs, as they require a somewhat different set of ethical expertise and on-the-ground knowledge of risks. Of course, scaling the ESR to other granting organizations will not include research that does not seek funding. This is a trade-off in the implementation of the ESR: By using funding decisions as an incentive for ESR participation, we alleviate self-selection concerns inherent in voluntary ethics and societal review processes but cannot reach those who do not require grant funding. At the same time, we hope that, if researchers begin engaging in this process through grant funding, it might help facilitate a culture shift to include the projects that do not require grant funding.

If the ESR continues to produce positive results, we hope to generalize it beyond our university. We believe that enabling other universities to stand up their own ESR will require easily adaptable materials (e.g., ESR statement prompt, panelist feedback forms), as well as workshops to support those interested in running their first ESR process. Growing beyond one university may make it more feasible for journals and conferences to consider ESR review as a requirement (12).

Ongoing evaluation of the ESR can help resolve outstanding research questions. What long-term impacts, if any, does an ESR have on the research projects or the community's reactions to them? Is an ESR more or less effective in certain fields and areas of research? Does the ESR have an impact on overall cultural attitudes toward ethics review among the researchers? Are certain aspects of the ESR process driving these changes over others? How should the ESR panel include stakeholders outside the university who represent different perspectives? Our methods in this study come with attendant limitations. The current evaluation may be subject to novelty effects, with researchers reacting more positively to the process due to it being different from their usual patterns. Ongoing evaluation of the program can help test long-term opinions and illuminate the pluses and minuses of this review process under different circumstances. Longitudinal study will also enable investigation of the long-term arc for those projects that went through the ESR review process: Did they produce better outcomes than projects that did not go through such review? Additionally, no project was ultimately denied funding; our data do not yet cover this case. Finally, this multimethod study was also not randomized, so we cannot and do not make any causal claims.

Conclusion

Research that has the risk of negatively effecting society, either immediately or through downstream applications, falls outside the jurisdiction of existing review procedures because many of these procedures exclude societal risks of harm. In this article, we introduce a process that operates in collaboration with a funding program to encourage ethical and societal reflection, only releasing grant funding when ethical and societal reflection is complete. Evaluation of the process over 1 y suggests that researchers found it valuable in broadening their ethical lenses and are willing to continue to submit to it despite the added commitment.

Materials and Methods

This section describes the ESR process in greater detail.

IRB Approval. This study, including the survey and interview consent process, was reviewed and approved by the IRB at Stanford University. On recruitment, participants were informed about the content of the study and the handling of the data. Opt-in consent procedures were used for both the survey and interview; only those who consented participated. Informed consent was also obtained from ESR panelists for analysis of their feedback. Only one panelist declined consent and their feedback was thus excluded from our analysis.

Funding Program. The most critical institutional feature of the ESR is the collaboration with a funding program. This collaboration enables completion of the ESR process to trigger release of funds. Funding is a rare moment of institutional leverage in the university: While most AI research proceeds without IRB review at our university, researchers are often in search of funding.

We partnered with a cross-departmental institute at our university that runs funding competitions with both 1) a large, multi-PI grant competition with a small number of projects receiving substantial funding and

2) a smaller seed-grant competition with many projects receiving less funding. Working with our team, the institute (program) added a requirement for an ESR statement for each grant submission. The program performed its typical merit review on all grant submissions and sent the ESR the proposals that they were interested in funding. The ESR then performed its own internal process on those proposals and reported its outcome and recommendation to the program.

ESR Statement. To aid researchers in structuring their thinking, the ESR statement prompt asks researchers to organize their statement into three parts. The full instructions are in *SI Appendix*. The first part articulates the risk: What happens when this research leaves the laboratory and becomes commercialized outside of your direct control, or when your study gets publicized and turned into public policy? The second part is a mitigation principle: What principle should researchers in the field follow to mitigate this risk in their work? Third is the specific research design: Describe how that mitigation principle is instantiated concretely in your proposed research design.

The instructions include examples. For example, if the first part includes a risk that a new healthcare algorithm is biased against Black members of society, a researcher might propose in the second part that all such algorithms must be audited against risks for underrepresented groups then describe how they will collect data to audit the algorithm against bias for Blacks, Latinx, Native American, and other underrepresented groups. In future iterations, based on researcher feedback, we also plan to include prompts suggesting common categories of issues that arise in ESR processes. The minimum ESR statement length ranges from one page to several pages depending on the project topic, size, and funding level.

Panel Review. The funding program next performs its grant merit review process and selects proposals that it would like to fund. The proposals and their ESR statements are then forwarded on to the ESR for feedback. The ESR's goal is not to filter out projects with any modicum of risk—instead, when possible, the goal is to aid the researchers in identifying appropriate mitigation plans.

The ESR faculty panel is composed to bring together diverse intellectual perspectives on society, ethics, and technology. Our panel thus far represents faculty from the humanities, social sciences, engineering, and medicine and life sciences. Their departments at our institution include Anthropology, Communication, Computer Science, History, Management Science & Engineering, Medicine, Philosophy, Political Science, and Sociology. Their interests include algorithms and society, gender, race, ethics, history of technology, collective action, medical anthropology, moral beliefs, medical ethics, social networks, AI, robotics, and human–computer interaction. Many other disciplines and identities can and should be included as well. Currently, the ESR panel is formed by the faculty ESR chairs, and faculty agree to continue on an annual basis. We did not study panelist motivations in our evaluation; however, informal discussions indicated that many panelists felt that ESR issues touched on their own research interests and that being asked to review a small handful of (three) proposals was not perceived as onerous. Two-thirds of the panel from the first year agreed to return for the second year.

Each proposal is assigned to at least two panel members, one representing the broad field of inquiry of the proposal (e.g., medicine, engineering, social science) and one representing a complementary perspective. A few panelists take on the role of chairs in facilitating the feedback process, overseeing the feedback process for individual proposals. To help with training, panelists are provided with example past proposals and the ESR responses for them. The ESR panel then meets synchronously to discuss particularly controversial or challenging projects.

Iteration and Approval. All researchers receive the free-text feedback provided by the two panelists. A subset are told that the ESR has completed its process on the projects and it will recommend the project for funding, though it welcomes further discussion if the researchers desire. Typically, these projects have low levels of concern from the panel.

The second subset of proposals decided upon by the ESR committee are asked to respond to the ESR's feedback. The ESR chairs make themselves available for conversation and consultation with the researchers. When the researchers respond, the response is passed back to the relevant panelists, who provide their thoughts and recommendation to the ESR chairs. The ESR chairs then draft a response to the researchers representing the ESR's thoughts and their own assessment and send it back to the researchers. Future iterations remain on email if the discussion is converging or can switch to a synchronous meeting if not. The ESR chairs become the point of contact for the researchers following the first round of feedback in order to avoid jeopardizing colleague relations (the ERB first-round feedback is authored anonymously) and to help facilitate the most challenging projects.

Data Availability. Anonymized survey, interview, and content analysis data have been deposited in the Open Science Framework: https://osf.io/mv4p6/, https://osf.io/vpq9m/, and https://osf.io/gk2j3/ (42–44). Due to the small number of observations and identifiability of respondents in the raw interview and content analysis data, only the coded data are made available.

ACKNOWLEDGMENTS. We thank the Stanford Institute for Human-Centered Artificial Intelligence for their collaboration. This work was supported by the Public Interest Technology University Network; Stanford's Ethics, Science, and Technology Hub; Stanford's Institute for Human-Centered Artificial Intelligence; and NSF Grant ER2-2124734. We also thank Ashlyn Jaeger, James Landay, Fei-Fei Li, John Etchemendy, Deep Ganguli, and Vanessa Parli for their support; the faculty panelists on the ESR for their time, insight, and energy; the researchers who engaged with the ESR for their effort and feedback; and Adam Bailey at Stanford's Institutional Review Board for his advice and Mary Gray at Microsoft Research for her guidance.

1. R. Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (John Wiley & Sons, 2019).
2. H. W. Rittel, M. M. Webber, Dilemmas in a general theory of planning. *Policy Sci.* **4**, 155–169 (1973).
3. R. McDermott, P. K. Hatemi, Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30014–30021 (2020).
4. Department of Health, Education, and Welfare, "The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research" (Department of Health, Education, and Welfare, 1979).
5. United States Department of Health and Human Services, Common rule. *Code Fed. Regul*. Title 45, §46.111 (2018).
6. J. Metcalf, K. Crawford, Where are human subjects in big data research? The emerging ethics divide. *Big Data Soc.* **3**, 2053951716650211 (2016).
7. M. Gray, D. J. Watts, E. Horvitz, Microsoft Research Ethics Review Program & IRB. https://www.microsoft.com/en-us/research/microsoft-research-ethics-review-program-irb/. Accessed 1 September 2021.
8. M. Angell, Medical research on humans: Making it ethical. *New York Rev*. (2015). https://www.nybooks.com/articles/2015/12/03/medical-research-humans-making-it-ethical/. Accessed 1 September 2021.
9. R. A. Calvo, D. Peters, AI surveillance studies need ethics review. *Nature* **557**, 31–32 (2018).
10. S. R. Jordan, *Designing an artificial intelligence research review committee* (2019). https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf. Accessed 1 September 2021.
11. G. Falco *et al.*, Governing AI safety through independent audits. *Nat. Mach. Intell.* **3**, 566–571 (2021).
12. G. Abuhamad, C. Rheault, Like a researcher stating broader impact for the very first time. arXiv [Preprint] (2020). https://arxiv.org/abs/2011.13032 (Accessed 4 August 2021).
13. D. K. Plemmons *et al.*, A randomized trial of a lab-embedded discourse intervention to improve research ethics. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1389–1394 (2020).
14. J. Petelka, K. Shilton, M. Finn, *Writing security: A curriculum intervention for computer security ethics* (2021). hdl.handle.net/2142/109700. Accessed 1 September 2021.
15. J. Borenstein, A. Howard, Emerging challenges in AI and the need for AI ethics education. *AI Ethics* **1**, 61–65 (2021).
16. J. S. Saltz, N. I. Dewar, R. Heckman, *Key Concepts for a Data Science Ethics Curriculum* (Association for Computing Machinery, 2018).
17. K. Cronin-Furman, M. Lake, Ethics abroad: Fieldwork in fragile and violent contexts. *PS Polit. Sci. Polit.* **51**, 607–614 (2018).
18. L. A. Fujii, Research ethics 101: Dilemmas and responsibilities. *PS Polit. Sci. Polit.* **45**, 717–723 (2012).
19. E. Montague *et al.*, The case for information fiduciaries: The implementation of a data ethics checklist at Seattle Children's Hospital. *J. Am. Med. Inform. Assoc.* **28**, 650–652 (2021).
20. American Political Science Association, *Principles and guidance for human subjects research* (2020). https://www.apsanet.org/Portals/54/diversity%20and%20inclusion%20prgms/Ethics/Final_Principles%20with%20Guidance%20with%20intro.pdf?ver=2020-04-20-211740-153. Accessed 1 September 2021.
21. American Psychological Association, *Ethical principles of psychologists and code of conduct* (2017). https://www.apa.org/ethics/code. Accessed 1 September 2021.
22. American Sociological Association, *Code of ethics* (2018). https://www.asanet.org/sites/default/files/asa_code_of_ethics-june2018a.pdf. Accessed 1 September 2021.
23. N. Bliss *et al.*, An agenda for disinformation research. arXiv [Preprint] (2020). https://arxiv.org/abs/2012.08572 (Accessed 4 August 2021).
24. K. Hartmann, K. Giles, "The next generation of cyber-enabled information warfare" in *2020 12th International Conference on Cyber Conflict (CyCon)* (IEEE, 2020), vol. 1300, pp. 233–250.

25. R. Zellers *et al.*, "Defending against neural fake news" in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2019), vol. 32, pp. 9054–9065.
26. A. Alkhatib, M. Bernstein, "Street-level algorithms: A theory at the gaps between policy and decisions" in *CHI Conference on Human Factors in Computing Systems Proceedings* (Association for Computing Machinery, 2019), pp. 1–13.
27. B. McInnis, D. Cosley, C. Nam, G. Leshed, "Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk" in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2016), pp. 2271–2282.
28. M. L. Gray, S. Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Eamon Dolan Books, 2019).
29. M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* **5**, 2053951718756684 (2018).
30. J. Buolamwini, T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification" in *Proceedings of Machine Learning Research* (MLResearchPress, 2018), vol. 81, pp. 77–91.
31. D. Danks, A. J. London, "Algorithmic bias in autonomous systems" in *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (IJCAI, Inc., 2017), vol. 17, pp. 4691–4697.
32. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
33. T. Gebru *et al.*, "Datasheets for datasets" in *Proceedings of the 5th Workshop on Fairness, Accountability, and. Transparency in Machine Learning* (2018), vol. 5, pp. 86–92.
34. H. Zhu, B. Yu, A. Halfaker, L. Terveen, "Value-sensitive algorithm design: Method, case study, and lessons" in *Proceedings of the ACM on Human-Computer Interaction* (2018), vol. 2, 1–23.
35. S. Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need* (MIT Press, 2020).
36. L. Winner, Do artifacts have politics? *Daedalus* **109**, 121–136 (1980).
37. R. K. Merton, The unanticipated consequences of purposive social action. *Am. Sociol. Rev.* **1**, 894–904 (1936).
38. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016).
39. S. U. Noble, *Algorithms of Oppression* (New York University Press, 2018).
40. V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).
41. S. Wachter-Boettcher, *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* (W. W. Norton & Company, 2017).
42. M. S. Bernstein *et al.*, ESR Content Analysis Data 2020. Open Science Framework. https://osf.io/mv4p6/. Deposited 6 December 2021.
43. M. S. Bernstein *et al.*, ESR Survey Data 2020. Open Science Framework. https://osf.io/vpq9m/. Deposited 6 December 2021.
44. M. S. Bernstein *et al.*, ESR Interview Data 2020. Open Science Framework. https://osf.io/gk2j3/. Deposited 6 December 2021.