PNAS
www.pnas.org

Supplementary Information for

Ethics and Society Review: Ethics Reflection As A Precondition to Research Funding

Michael S. Bernstein, Margaret Levi, David Magnus, Betsy Rajala, Debra Satz, Charla Waeiss

Margaret Levi.
E-mail: mlevi@stanford.com

This PDF file includes:

Supplementary text
Tables S1 to S4 (not allowed for Brief Reports)

Supporting Information Text

Materials and Methods

Materials. This section includes the full text of the original 2020 ESR statement prompt (relevant to the evaluation described in the main paper), the updated 2021 ESR statement prompt (used in the 2021 ESR process), the survey and interview instruments administered to 2020 ESR participants, and the 2021 staff panel calibration and review process. At the time of administration, the ESR was known as the Ethics Review Board (ERB). We also include a description of the staff panel calibration process we employed for the 2021 ESR.

For those interested in setting up their own ESR, we recommend referring to our 2021 ESR statement prompt, as it incorporates a number of revisions we identified in our 2020 evaluation of the process.

2020 ESR statement prompt. For any questions, contact the ERB chairs at [email address].

- Michael Bernstein, Associate Professor of Computer Science
- Margaret Levi, Sara Miller McCune Director of the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford, Professor of Political Science, and Senior Fellow at the Woods Institute for the Environment
- David Magnus, Director, Stanford Center for Biomedical Ethics, Thomas A. Raffin Professor of Medicine and Biomedical Ethics and Professor of Pediatrics, Medicine and By Courtesy of Engineering
- Debra Satz, Marta Sutton Weeks Professor of Ethics in Society, and Vernon R. and Lysbeth Warren Anderson Dean of the School of H&S

Detail the ethical challenges and possible negative societal impacts of the proposed research. How will you mitigate them? For example, consider autonomy and consent from those you are getting data from, who is and is not represented in the training and test data, and whether your approach is importing bias from existing machine learning models.

We suggest no more than one page as a starting point, as a supplement to your grant proposal. The ERB panel will read both the grant proposal and the ERB statement. We of course do not expect that all of the ethical considerations for your projectcan be described in one page. The ERB statement kickstarts an iterative process, and the ERB may ask for responses in reaction to what you wrote. If you need more space in the initial statement, email [email address] (James Landay, HAI Seed Grants) and [email address] (Michael Bernstein, ERB).

Organize your statement into two parts:

1. *Describe the ethical challenges and societal risks.* What are the most important challenges you face with your project? Consider the following three groups in your response: (1) Society: the society targeted by the research, considered as a whole (e.g., American society); (2) Subgroups within society: risks are not distributed equally amongst a society, and marginalized subgroups may be especially vulnerable (e.g., LGBTQ+ individuals); (3) Global: impacts on the world as a whole, or on societies that are not directly targeted by the research but that may be impacted by the research (e.g., potential abuse in developing regions).

2. *Articulate general principles that researchers in your field should use to eliminate or mitigate these issues, and translate those principles into specific design decisions you are making in your research.* Think of what happens when someone else builds on your work. What principles should others in this field follow when faced with similar tradeoffs? How does your proposal instantiate those general principles? If your research team does not currently have required expertise or perspectives represented, how will you obtain them?

A brief example: If (1) includes a risk that a new healthcare algorithm is biased against Black members of society, you might propose in (2) that all such algorithms must be audited against risks for under-represented groups, then describe how you will collect data to audit the algorithm against bias for Blacks, Latinx, Native American, and other under-represented groups.

Why are we doing this?

AI research is now routinely criticized by academics and by the public for its accelerating negative impacts on society. Today's academic procedures for ensuring ethical research, derived from medicine as described in the Belmont Report and encoded in requirements such as the Institutional Review Boards (IRB), do not provide guardrails to address these critiques. Specifically, IRBs are designed to evaluate harms to individuals (research participants) rather than harms to society. HAI has thus provisionally created an Ethics Review Board (ERB), focused on evaluating the benefits and harms to society of funded HAI research, and helping guide researchers when needed. Following the academic merit review process for any HAI research grant, the ERB reviews accepted proposals to assess foreseeable benefits and risks to society, to subgroups within society, and to the world, before the grant is funded.

How will this work?

HAI will first conduct its academic merit review on the proposals. Once it decides which ones it would like to fund, HAI will forward the proposals and their accompanying ERB statements to the ERB. A panel of ERB members will read the statements alongside the original grant. Most typically, the ERB will send written feedback and request a response or revision to the project's ERB statement. The ERB can also help connect projects to collaborators or stakeholders if needed or requested. The

ERB's goal is to help guide the conversation, and bring in experts to help expand the horizon of foreseeable harms and how to mitigate them. If a case does arise where the PIs and ERB cannot align on an approach, the case will be turned over to HAI executive leadership for a final decision. The goal of the ERB is not to act as a filter: it is to work with PIs to ensure successful and pro-social research outcomes.

Please direct any questions to [email address].

2021 ESR statement prompt. For any questions, contact the ESR chairs at [email address]:

- Michael Bernstein, Associate Professor of Computer Science
- Margaret Levi, Sara Miller McCune Director of the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford, Professor of Political Science, and Senior Fellow at the Woods Institute for the Environment
- David Magnus, Director, Stanford Center for Biomedical Ethics, Thomas A. Raffin Professor of Medicine and Biomedical Ethics and Professor of Pediatrics, Medicine and By Courtesy of Engineering
- Debra Satz, Marta Sutton Weeks Professor of Ethics in Society, and Vernon R. and Lysbeth Warren Anderson Dean of the School of H&S

In this document:

- What goes in the ESR statement?
- What are common risks and mitigations included in ESR statements?
- The ESR is not the IRB, and focuses on different issues
- Example ESR statements
- Why are we doing this?
- What's the process?

What goes in the ESR statement?

Describe the ethical challenges and posisble negative societal risks of the proposed research, and how you will mitigation them. We strongly suggest the following organization for each risk:

- *Description*: what is the risk? Think about what happens when this research leaves the lab and becomes commercialized outside of your direct control, or when your study gets publicized and turned into public policy. (e.g., "The algorithm may be used to discriminate against low-income students.")
- *Mitigation principle*: what principle should researchers in your field follow to mitigate this risk in their work? (e.g., "We follow a principle that public policy algorithms should be audited against minoritized groups prior to publishing, and that audit be included in the research article.")
- *Research design*: describe how that mitigation principle is instantiated concretely in your proposed research design. What commitments are you making? (e.g., "We will implement our sensing algorithm locally on the user's device, and advocate for this privacy approach in papers and public talks about this work.")

We suggest no more than one page as a starting point, as a supplement to your grant proposal. The ESR panel will read both the grant proposal and the ESR statement. We of course do not expect that all of the ethical considerations for your project can be described in one page. The ESR statement kickstarts an iterative process, and the ESR may ask for additional conversation in reaction to what you wrote. If you need more space in the initial statement, email [email address] (James Landay, HAI Seed Grants) and [email address] (Michael Bernstein, ESR).

What are common risks and mitigations included in ESR statements?

By analyzing previous projects and ESR responses, we have identified the most common set of topics that researchers and the ESR raise. We suggest that you think about whether each of these categories are salient risks for your project: [See S1 for more detail].

Risk	Example Principle	Example Mitigation
Representativeness Insufficient or unequal representation of data, participants, or intended user population Example: data collection process for a wellbeing sensing algorithm would undersample low-income populations	Algorithm training data and evaluation should include communities likely to be impacted by the algorithm	Commitment to explicitly recruit low-income individuals to ensure that their data is included in the training, and that their voices are heard in the evaluation
Diverse design and deployment Incorporating relevant stakeholders and diverse perspectives in the project design and deployment process Example: an algorithm for fairer school choice not considering the voice of those historically disadvantaged by school choice mechanisms	Algorithms for social choice should directly consult with stakeholders who would be impacted by their deployment	Commitment to include a PI on the project who brings expertise on experiences in education from historically disadvantaged groups Commitment that the researchers will engage in stakeholder discussions or participatory design processes with members of historically disadvantaged groups
Dual use The technology being co-opted for nefarious purposes or by motivated actors Example: algorithmic sensing advances might be co-opted by authoritarian governments or employers for surveillance	Sensing algorithms should place control in the hands of those being sensed	Commitment to develop an architecture where the sensing algorithm operates on the user's device and keeps all data local Commitment to use the "bully pulpit" of Stanford researchers to describe the importance of this architecture in papers and talks about the research
Harms to subgroups Harms to populations that could arise following from the research's success or translation into policy Example: teacher job loss due to better education algorithms	Educational interventions should be designed as amplifying teachers' abilities, rather than replacing teachers	Commitment to designing the algorithm in a way that requires teacher input and oversight

Table S1. Common risks and mitigation strategies raised by researchers and the ESR in the 2020 ESR process.

The ESR is not the IRB, and focuses on different issues

Institutional Review Boards (IRBs) are prohibited from considering ethical and societal risks that impact human society rather than human subjects. As the U.S. Common Rule (§46.111) states, "The IRB should not consider possible long-range effects of applying knowledge gained in the research (e.g., the possible effects of the research on public policy) as among those research risks that fall within the purview of its responsibility." The ESR exists because much AI research does not directly involve human subjects, and thus is outside of IRB purview, but does impact human society.

Do not discuss issues that should be in IRB scope in your ESR statement: those issues will be reviewed by the IRB when you submit your human subjects protocol. Any risk directly impacting participants in your research, such as data privacy, physical harms, or fair wages for participants in your studies, is not relevant to the ESR. In contrast, the ESR is interested in privacy, harms, and wages that will arise after this research leaves the lab. [See S2 for more detail.]

	IRB: Institutional Review Board	ESR: Ethics and Society Review
Focus	Risks to human subjects	Risks to human society
Time	Risks arising during the research (e.g., during the study)	Risks arising after the research is complete (e.g., during wider deployment or commercializations, in public policy)
Example risks	Privacy for participants Impacts on the study population during the study Participant payment	Privacy for those using the algorithm in industry or civil society Impacts on marginalized groups after deployment Impacts on wages and jobs

Table S2. Summary of the differences between the ESR and IRB.

Example ESR statements

[*With permission from PIs, the ESR provided access to example ESR statements from the 2020 ESR review. Access to these examples was restricted internally to Stanford University.*]

Why are we doing this?

Artificial intelligence (AI) research is routinely criticized for its negative impacts on society. We lack adequate institutional responses to this responsibility: AI research often falls outside the purview of existing research mechanisms such as the Institutional Review Board (IRB), which are designed to evaluate harms to human subjects rather than harms to human society. In response, we have developed Ethics and Society Review (ESR), a feedback panel that works with researchers to mitigate negative ethical and societal aspects of AI research. The ESR serves as a requirement for funding: researchers cannot receive grant funding from HAI until they complete the ESR process for the proposal. We have run the ESR process across over 40 proposals so far.

What's the process?

HAI will first conduct its academic merit review on the proposals. Once it decides which ones it would like to fund, HAI will forward the proposals and their accompanying ESR statements to the ESR. A panel of ESR members will read the statements alongside the original grant. The ESR may send written feedback or schedule a conversation. The ESR can also help connect projects to collaborators or stakeholders if needed or requested. The ESR's goal is to help guide the conversation, and bring in experts to help expand the horizon of foreseeable harms and how to mitigate them. If a case does arise where the PIs and ESR cannot align on an approach, the case will be turned over to HAI executive leadership for a final decision. The goal of the ESR is to act as a coach, not a reviewer.

Please direct any questions to the [email address].

Case studies: detail. **AI Teaching Agents: displacement, representation, and optimization targets.** One of the larger grants reviewed by the ESR, which included six faculty crossing Computer Science, Education, and Psychology, focused on the creation of a new generation of AI tutoring systems. The systems would be evaluated through field deployments online and in brick-and-mortar schools.

The ESR focused its initial feedback on three questions. First, the panel wanted to hear more about whether the research would be used as a justification to displace teachers' jobs (e.g., dual-use concerns). In their ESR statement, the researchers had stated, "we design our tools to force multiply teachers so that they can have more impact, and to only provide fully autonomous education in contexts where there is no human option". The panelists asked the researchers to clarify what, concretely, in the proposed system would make it more useful as a force multiplier than as a replacement.

The second discussion point was: who will be represented in the data? It would be important to make sure that the resulting model is not just useful to the communities already have the capacity to benefit. The panel asked the researchers to clarify whether background diversity would be reflected in the dataset collection, and how that diversity would be recruited (e.g., representativeness concerns).

The final question was about what the models would learn: depending on how it is implemented, optimizing for long-term engagement could cause the algorithm to increase its reward most effectively by focusing on the learners who it is most likely to be able to retain. This would mean it would bias against groups of learners who drop out at high rates — traditionally groups of low socio-economic status students — since the policy would learn that they are unlikely to be retained (e.g., harms to subgroups concerns), so it should focus its effort on others. The panelists asked the researchers to propose how they would guard against this?

The researchers then responded to each of the three points. With respect to job displacement, the researchers pointed to research identifying teachers as the least likely jobs to be displaced by AI, and committed to a specific design constraint on their systems that they assume a teacher copresent in the room alongside the student. With respect to representation, the team committed to the process being led by one of their team members who specializes in the design of inclusive educational experiences for marginalized communities, and further committed to explicit transfer testing with diverse groups. Finally, with respect to undesired outcomes of the optimization criteria, the team stated that answering those questions was an explicit goal of the research, and committed to engaging with experts in educational equity throughout the project.

The panelists felt that the responses responded to the questions and committed to specific mitigation strategies, and so passed the proposal.

Remote sensing AI: malicious actors. A seed grant featuring faculty from Earth Systems Science and Computer Science proposed an approach to creating remote sensing models to be used in sustainability applications. In their ESR statement, the researchers pointed out two possible threats: that the models could be adopted by intelligence and military organizations in authoritarian states to enhance mass surveillance (e.g., dual-use concerns), and that the representations they developed might perform differently in different parts of the world (e.g., diverse design and deployment concerns). The researchers pointed out that the mass surveillance applications were less likely to arise from their research because such state actors already had resources to collect the large-scale data needed to train such models, making the incremental risk of this project more minor; and the researchers committed to focusing their models solely on sustainability applications. With respect to performance imbalances, the team committed to benchmarking their models globally, and specifically focusing on Africa, to challenge the status quo of similar models focusing on America.

The ESR panelists, in their responses, focused on a theme of how to guarantee non-malicious use. They pointed out that naming the possibility of malicious use itself does not carry teeth, and that the statement relies on an assumption that other datasets are more attractive to malicious actors than the datasets and models produced in this research.

The panelists felt that a paired forensic model was a good commitment, and suggested that the researchers gather data on how people in each area are surveyed currently and what impact this shift might give rise to. Ultimately, the panelists felt that this feedback was sufficient and did not request the panelists to respond.

Stress sensing: privacy. One of the funded seed grants, proposed by faculty in Medicine and Electrical Engineering, focused on noninvasive stress sensing at work. In their ESR statement, the researchers expressed an awareness of possible concerns about repressive governments and oppressive employers as two examples of groups that would use this technology to surveil and potentially harm (e.g., dual-use concerns). The statement did not provide principles that researchers should use to mitigate that risk, or concrete instantiations of those principles in their research, instead stating that the team would take user agency and privacy seriously.

The panelists asked the researchers to iterate on their statement and explain specific design features that they would be including to address the risks. In addition, one panelist pointed the researchers to prior work on commiting to returning information about stress levels to the participants at the conclusion of the study and on an ongoing basis if appropriate.

The researchers and cognizant ESR chair in charge of that proposal then met to discuss the feedback. Following the discussion, the researchers wrote their response to the committee feedback, which focused on how the researchers would protect data (e.g., privacy concerns), but did not discuss what should be done early in the research to account for risks once the technology leaves the lab and is commercialized.

In response, the researchers committed to explaining the privacy-preserving aspects of their design, and the importance of those aspects, in all papers arising from this research — essentially using their bully pulpit at our institution to push for norms in how this technology is developed and deployed. The ESR felt that the combination of this commitment with the conversations held with the researchers was appropriate to the level of risk, and agreed to recommend the proposal.

Survey instrument. The survey was administered from November 10, 2020 to February 11, 2021 via Qualtrics to the lead researchers on all seed grants that participated in the ESR.

1. Which Stanford department are you primarily affiliated with?
2. What is the title of the HAI proposal that you submitted?
3. Not including the ERB, how long ago was the last written or spoken conversation that you had about ethics in one of your research projects?
    - Within 1 day
    - Within 1 week
    - Within 1 month
    - Within 6 months
    - Within 1 year
    - Longer than 1 year
    - Cannot recall
4. Prior to the proposal submitted for the HAI's seed grant, had any of your projects been reviewed by the Ethics Review Board (ERB), or any other ethics review process? *Please recall that IRBs do not perform ethics review — they are focused only on risks to research participants, not risks to society.*
    - Yes, by Stanford's ERB
    - Yes, by another organization
    - No, I haven't had any other projects reviewed by an ethics committee

    »If [Yes, by another organization] is selected show question 5. If any other option is selected move on to question 6.
5. Please provide the name of the organization that conducted the ethics review of your previous project.
6. How strongly do you agree with the statement: The ERB process influenced the design of my research project.
    - Strongly agree
    - Agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Disagree
    - Strongly disagree
7. Which aspect of the ERB process did you find most helpful?
    - Writing the ethics statement as part of your proposal.
    - Receiving feedback from the ERB on your ethics statement.
    - Revising your original ethics statement based on ERB feedback.
    - Using the ERB as a resource to connect with other experts or stakeholders.
    - Other [freetext]
8. Which aspect of the ERB process did you find least helpful?
    - Writing the ethics statement as part of your proposal.
    - Receiving feedback from the ERB on your ethics statement.
    - Revising your original ethics statement based on ERB feedback.
    - Using the ERB as a resource to connect with other experts or stakeholders.
    - Other [freetext]

9. Research projects occasionally garner public criticism (e.g., in news media or social media) for ethical missteps. Examples in the last few years at Stanford included election experiments, engineering innovations that encourage worker displacement, and dataset bias. How strongly do you agree with the claim: ERB review helped mitigate the possibility of public criticism of my project.

    - Strongly agree
    - Agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Disagree
    - Strongly disagree

10. Would you be willing to submit a future research project to the ERB for review?

    - Yes, I would voluntarily submit a future project to the ERB for review.
    - Yes, if required by a funding organization or the university.
    - No, there are no circumstances where I would submit a future project to the ERB for review.

11. The ERB's goal is to help work with PIs, understand tradeoffs, and brainstorm ways to mitigate negative outcomes. However, it may be called upon to render an opinion on an especially ethically problematic proposal. How strongly do you agree with the claim: the ERB should be empowered to reject an especially ethically problematic proposal.

    - Strongly agree
    - Agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Disagree
    - Strongly disagree

Semi-structured interview protocol. *First, we would like to ask you some questions about your experience considering the ethical implications of your work and your ethics training.*

1. How much do you typically engage with ethical issues in the research you do?
2. How much do researchers in your field take into account the ethical implications of their research?
3. Do you believe that most faculty are sufficiently trained to consider the ethical issues with their research?
4. Graduate students?

   Next, we would like to ask you about your experience with the ERB process.

5. Through the ERB process, did you become aware of ethical issues you hadn't previously considered?
6. Why (why not) was the ERB able to make you aware of these issues? / Which aspects of the ERB process alerted you to these issues?
7. How helpful to you was the identification of these issues?
8. Did it lead to concrete changes in your project design?
9. What would you change about the ERB process?

   Now we would like to follow-up on a few questions we asked in the survey you completed after the ERB review process.

10. We had asked you, "Which aspect of the ERB process did you find least helpful?", you responded that [participant's response] was the least helpful. Why was that the least helpful aspect of the process for you?
11. We had asked you, "Which aspect of the ERB process did you find most helpful?", you responded that [participant's response] was the most helpful. Why was that the most helpful aspect of the process for you?
12. Research projects occasionally garner public criticism (e.g., in news media or social media) for ethical missteps. Examples in the last few years at Stanford included election experiments, engineering innovations that encourage worker displacement, and dataset bias. In our survey, we asked you, "How strongly do you agree with the claim: ERB review helped mitigate the possibility of public criticism of my project." You responded that [participant's response]. Why do you [participant's response]?

    Finally, we would like to get your thoughts on the authority the ERB should have and your willingness to engage in ethical reviews in the future.

13. The ERB's main purpose is to work with researchers to reflect on and design ethics into their research. How would you say your experience aligns with the intended purpose of the ERB? (*Question added to interview protocols used after December 24, 2020.*)
14. The ERB's main purpose is to work with researchers to reflect on and design ethics into their research. However, the ERB is also called on to make decisions about when to decline to recommend a project for funding due to ethical concerns. In our survey, we asked you, "How strongly do you agree with the claim: the ERB should be empowered to reject an especially ethically problematic proposal." You responded that [participant's response]. Why do you [participant's response]?
15. Under what conditions should the ERB not recommend a project for funding, if ever?
16. Are there any circumstances where the ERB should recommend a project for funding even though considerable potential harms have been identified?
17. Under what conditions do you think projects should be reviewed by an ethics review panel?
18. Should the ERB make summaries of their recommendations publicly available?
19. Under what conditions would you be willing to submit a future research project to the ERB for review?
20. Is there anything else you would like to share about your experience with the ERB?

2021 staff panel calibration and review process. Four PhD-level staff conducted the initial triage of the 2021 ESR proposals. Based on the 2020 ESR and the distribution of risk assessment across the proposals, we intended for the staff panel to identify approximately one-third to one-half of proposals that would not need to undergo faculty review. Each panelist participated in a calibration exercise as follows:

A representative sample of eight proposals was drawn from the 2020 ESR proposals. The staff panelists were provided with a suggestive rubric to accompany their review of the proposals. This rubric posed the following questions:

- Did the statement address specific ethical concerns and societal implications related to the PIs' project?
  - Which ethical issues did they discuss? [*Table 2 from the main paper was provided as reference.*]
  - Did the statement identify mitigation principles to these concerns?
  - Which mitigation principles did they discuss?
- Did the statement discuss concrete applications of these mitigation principles in their research design?
  - Were the mitigation strategies reasonably appropriate for the concerns PIs discussed?
  - Are any obvious ethical concerns and mitigation strategies missing?

In the calibration round, staff panelists were asked to provide a summative score (1-5) for each proposal. This scoring matched the guidance provided to panelists in the 2020 ESR round, where 1 = the panelist had no concern for the ethical and societal implications of the projects and 5 = very concerned. In all but one case, the average of staff panelists' assessment was within less than one point of the 2020 faculty panelists' assessment. For the remaining outlier, which staff scored considerably higher than faculty, the staff panelists discussed the project with the ESR chairs to identify the source of the misalignment: the faculty panel had additional knowledge of the researchers' previous work that provided reasonable mitigation strategies for risks identified in their ESR statement.

Following this calibration, the staff panelists participated in the initial round of the 2021 ESR review process. Two staff panelists were assigned to each proposal and asked to determine whether, based on the project proposal and ESR statement, the proposal warranted faculty review. If at least one staff panelist recommended faculty review of a proposal, the project was forwarded to the faculty panel.

Methods. This section provides further information on the data collection and analysis. It also includes descriptive information on the survey and interview sample populations, and additional evidence from the content analysis and interviews that supplement the findings included in the main paper.

Following researchers' participation in the ESR process, we invited the lead researchers from the 35 seed grants to complete a survey about their ESR experience. The survey instrument was administered via Qualtrics from November 10, 2020 to February 11, 2021. Twenty-six researchers representing twenty-five proposals responded to the survey, a 71% project-level response rate (see Table S3 for a breakdown of proposals, iteration, survey respondents and interview respondents by field.)

We also invited the lead researchers to each participate in a semi-structured interview, which took place from December 16, 2020 to February 5, 2021. To avoid any overt conflicts, the coauthor who conducted the interviews was neither an ESR chair nor panelist. The interviews were conducted virtually, with sixteen researchers representing fifteen different proposals participating, a 43% project-level response rate. As the interview protocol illustrates, researchers were asked about their experience engaging with ethical issues, their perceptions about ethical considerations by faculty and graduate students, whether and how the ESR process influenced their own ethical reflections and design of their research, the authority they believed the ESR should have, and their willingness to participate in the ESR process again.

Thirteen of the fifteen projects (87%) represented in the interviews were also represented in the survey. Therefore we are cognizant of factors conditioning researchers' selection into participating in the survey and interview that could also influence their reported perception of the ESR process. In particular, those who provided their feedback on the ESR process may have reacted more positively to it than those who did not, increasing their willingness to commit additional time to the survey and/or interview. At the same time, participants were candid in the interviews about the aspects of the ESR process that could use improvement and instances where the ESR did not achieve its goals. One participant even indicated that they agreed to an interview so that they could share the aspects of the ESR that did not seem to work for them.

In addition to soliciting researchers' feedback on the ESR process in our evaluation, one coauthor also conducted open coding of the ESR statements and panelist feedback. Using a grounded theory approach, they identified the common risk themes that participants and panelists noted in their statements and feedback, respectively. In the initial review, they organized the risks along the following themes: accountability, collaborator concerns, consent, dual-use, diverse deployment and design, general harm to others, human error, publicness, privacy, representativeness, research transparency, tool error, and other. Following a review with other coauthors, we revised the themes as follows: (1) we collapsed participant-specific consent and privacy concerns into an "IRB purview" category, as these risks are reviewed by the IRB and not the ESR; (2) we disaggregated the general harm to others category into "harms to society" and "harms to subgroups"; (3) we combined the human error and tool error categories together; and (4) we added a category for "methods and merit" to denote concerns that should fall under a traditional methods and merit review rather than the ESR review. After this reorganization and the subsequent recoding, a

second coauthor independently coded a random sample of 20 (57%) ESR statements and corresponding feedback. As noted in the main text, inter-rater reliability via Cohen's kappa averaged 0.96 per theme, with a range of 0.83 - 1.

On the whole, the ESR statement and panelists' feedback revealed that both PIs and panelists were grappling with many ethical issues that would not have otherwise been addressed. It also appears that panelists broadened the ethical scope for PIs in a responsive manner. For example, as a majority of panelists already discussed issues of representativeness in their ESR statements, panelists were less likely to raise those issues. Instead, panelists raised additional issues in those cases, like harms to subgroups and dual-use concerns. Of the 35 projects reviewed during this cycle, panelists on 26 of those projects (74%) raised new ethical issues that the PIs had not discussed in the ESR statements. Among the 25 projects that did not require iteration with the ESR, panelists raised new ethical issues for 18 of them (72%). This result illustrates that panelists' feedback was not constrained by the content of the ESR statements and served to broaden the ethical scope that PIs used in their projects.

Panelists pointed out ethical issues that researchers hadn't discussed in their ESR statement for all but 2 proposals that iterated with the ESR (80%). For those two proposals, panelists' feedback concentrated on requesting further details from researchers and offering mitigation strategies for some of the ethical issues that were unresolved by the discussion in the ESR statement. Only for 4 projects did panelists constrain their feedback to what was included in the ESR statement.

As Table S4 illustrates, in addition to raising new risks and deepening 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 mitigation strategies pertained to ethical issues the researchers raised in their ESR statement but left 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. While some panelists had no feedback, all projects received substantive feedback from at least one panelist. This lends some confidence to the ESR process and its goals; PIs encountered some feedback on their projects, even if it passed review without additional actions needed.

Field	Total seed grant proposals	Iterated proposals	Survey respondents	Interview respondents
Arts	2	0	1	0
Earth Science	2	0	2	2
Engineering	14	4	12	7
Humanities	1	0	1	0
Medicine	5	4	3	2
Social Science	11	2	6	4
Total	35	10 (29%)	25 (71%)	15 (43%)

Table S3. We determined the field assigned to each proposal based on the lead PI's own field. However, many of the proposals were submitted by interdisciplinary teams. Participants from each field responded to the survey. Participants from Earth Science, Engineering, Medicine, and Social Science participated in the interviews. Compared to the full seed grant population, perspectives from researchers in Engineering were over-represented in the interviews and we lacked responses from participants in the Arts and Humanities.

Panelist actions	Frequency
Identifying new risk	26
Deepending the conversation	22
No actions taken nor issues raised	9
Providing mitigation strategies	6
Recommending relevant research	4
Identifying a potential collaborator	2

Table S4. ESR panelists most commonly raised new risks that the researchers had not described in their statement, and deepened themes that the researchers had already raised.

As discussed in the main paper, some researchers reported that the act of writing the ESR statement and engaging with panelists' feedback led them to further discuss the ethical implications of their work and, in some cases, identify new issues they had not otherwise considered.

> I don't consider myself an expert, by all means, but I'm definitely very interested [in research ethics] and I think the [ESR] has started that spark in me. I'm very interested in doing it – like, actively pursuing it. But I still need to practice and learn...theory...it's necessary for these things, so I'm still in the learning phase, but I'm very motivated. Let's play. - Researcher, medicine

> I think just the brainstorming aspect of being like, "Oh yeah, there's an ERB, they're going to think about our project. We don't want to be embarrassed by not having thought about obvious downstream harms of our thing. So let's just sit down and spend two hours ... coming up with bad ways of using this technology." - Researcher, engineering

> It was great to have a conversation...the [feedback] I think was useful. And then in turn, being able to...go through your work...through your responses [to the feedback] with people. I think that was all very useful. I think it exposed me to certain things...like how my post-doc thinks...[that] I might not have known otherwise. - Researcher, engineering

> ...Having to write [the ESR statement] led to...the discussions...with the co-PI...that sort of brought [out additional issues]. [It's] a chain of events... But I think having to write [the statement]] was the necessary condition because otherwise [I] wouldn't have had that conversation [with my co-PI]. - Researcher, social science

We highlighted in the main paper our findings on whether and how the ESR affected the design of researchers' projects. In the interviews, four research teams identified significant changes that they made to their work, with one team even updating their research question to bring privacy concerns to the forefront of their project. For the two other respondents, one noted changes to how they engaged with local experts. The other modified their approach to privacy and attending to the dual-use nature of their project. Six other respondents were at such early points in the research process that, while they couldn't identify concrete changes the ESB process yielded, they anticipated their experience with the ESB had shaped and would continue to influence the development of their projects.

As noted in the main paper, participants expressed desire for the ESR to provide additional scaffolding for considering the ethical implications of one's research.

> I wonder if there was a way of walking through ... having "these are the kinds of considerations," kind of a little more granularity might help another ERB. Perhaps a training module or links to examples, so that one could either go through the training or review examples, etc., as part of completing the ERB. - Researcher, earth science

> The prompt was more like, "what do you think is wrong?" And...I don't think anything is accomplished. You're trying to figure out [how to] poke holes in your own story. It's very difficult. But ... if there was more specificity ... such as ... [specific questions like] "does your thing discriminate on a protected characteristic?" - Researcher, engineering