Ethics and Society Review (ESR) Statement

In this document:

[What goes in the ESR statement?](#_s6za119lq8hr)

[What are common risks and mitigations included in ESR statements?](#_yotz4qtjwi83)

[The ESR is not the IRB, and focuses on different issues](#_vi6oq5tfz4tr)

[Example ESR statements](#_96cx14fi4z6c)

[Why are we doing this?](#_ao1vzqk0wnd3)

## What goes in the ESR statement?

Describe the ethical challenges and possible negative societal risks of the proposed research, and how you will mitigate 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.")

## What are common risks and mitigations included in ESR statements?

The ESR has worked with over 70 proposals in collaboration with HAI. 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:

|  |  |  |
| --- | --- | --- |
| **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 consider 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 |

## 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.

|  |  |  |
| --- | --- | --- |
|  | **IRB** | **ESR** |
| 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 commercialization, in public policy) |
| Example risks | Privacy for participants  Impacts on 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 |

## 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.