HAI Seed Grants 2020:
Ethics Review Board (ERB) Statement

For any questions, contact the ERB chairs at ethicsreviewboard@lists.stanford.edu:

- 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 Bioengineering.
- 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 project can 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 landay@stanford.edu (James Landay, HAI Seed Grants) and msb@cs.stanford.edu (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., LGTBQ+ 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 ethicsreviewboard@lists.stanford.edu.