ESR: Ethics and Society Review of AI Research

Michael Bernstein, Computer Science
Margaret Levi, Political Science and CASBS
David Magnus, Medicine and Biomedical Ethics
Debra Satz, Philosophy and Dean of Humanities & Sciences

Stanford University


Defending Against Neural Fake News
Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi
Abstract
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
Joy Buolamwini
MIT Media Lab 75 Amherst St. Cambridge, MA 02139
JOYAB@MIT.EDU
Timnit Gebru
Microsoft Research 641 Avenue of the Americas, New York, NY 10011
TIMNIT.GEBRU@MICROSOFT.COM
Editors: Sorelle A. Friedler and Christo Wilson
Abstract
Word embeddings quantify 100 years of gender and ethnic stereotypes
Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou
PNAS
Original Research Article
Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management
Min Kyung Lee
Abstract
why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE
RUHA BENJAMIN
RACE AFTER TECHNOLOGY
"This book is downright scary—but...you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN
AUTOMATING INEQUALITY
HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR
VIRGINIA EUBANKS
How to Stop Silicon Valley from Building a New Global Underclass
GHOST WORK
Mary L. Gray and Siddharth Suri
RESEARCH
RESEARCH ARTICLE
ECONOMICS
Dissecting racial bias in an algorithm used to manage the health of populations
Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan
Datasheets for Datasets
TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

WE LACK INSTITUTIONAL RESPONSES TO AI ETHICS

Success requires that **everyone** participate, in the **formative** stages of research

- Opt-in approaches — office hours, design principles, and checklists [e.g., Madaio et al. 2020, Rakova et al. 2020, Mittelstadt 2019] — help those who self-select to participate
- Broader impacts in papers — e.g., NeurIPS and the FCA recommendation— are written after the research is complete

WHAT ABOUT THE IRB?

In the United States, IRB regulations focus on risks to **human subjects**, not risks to **human society**

"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**." [Common Rule 2018, §46.111]

So, most AI research currently falls outside IRB purview.

Sometimes, IRBs will take a broader lens, as in the Microsoft Research Ethics Review Program [Gray, Watts, and Horvitz 2013]

ESR: ETHICS AND SOCIETY REVIEW

[Bernstein et al. PNAS 2021]

An institutional process in collaboration with the Stanford Institute for Human-Centered Artificial Intelligence (HAI) that facilitates researchers in **mitigating negative ethical and societal aspects of AI research**

Designed as a **gate to access funding**: HAI grant funding is not released until the ESR process is completed for the grant

Grant application and ESR statement submitted to funding program → **Merit review by funding program** → **ESR triage** → **ESR panel** → **Recommendation to funding program**

Name the risks, articulate principles for mitigation, instantiate those principles in the research design

ESR panel ⇄ **Feedback & iteration**

Interdisciplinary panel, including Anthropology, Communication, CS, History, MS&E, Medicine, Philosophy, Political Science, and Sociology

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:

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

CASE STUDY: STRESS SENSING

FACULTY IN ENGINEERING & MEDICINE

Researchers **named concerns** surrounding surveillance by governments and employers, but stopped there

Panel feedback: what **specific research design** will mitigate these risks?

Meeting to discuss feedback

Description of **privacy-preserving architecture** and commitment to explain the importance of this architecture in **papers and public talks** about the work

THE ESR SO FAR

In collaboration with Stanford HAI and Woods Institute, the ESR has reviewed **92 proposals in its first two years**

In year one: all of the Hoffman-Yee grants and 29% of the seed grants iterated at least once with the ESR

So what happened, and what have we learned?

A BRIEF WORD ON OUR METHOD

Survey of lead researchers on all funded HAI seed grants

- 23/35 projects = 66% response rate

Follow-up **semi-structured interviews** with lead researcher

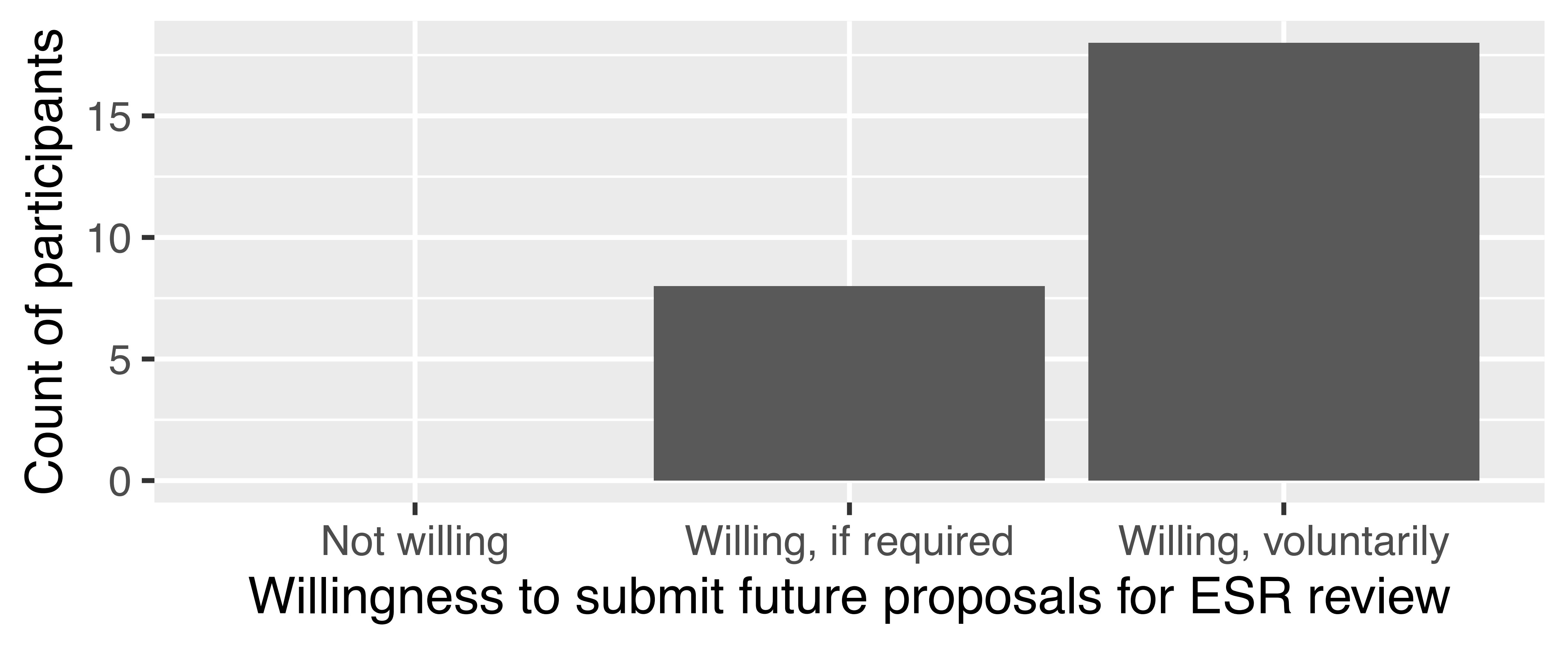
- 13 projects: Focuses in engineering (7), social science (4), earth science (2), and medicine (2)

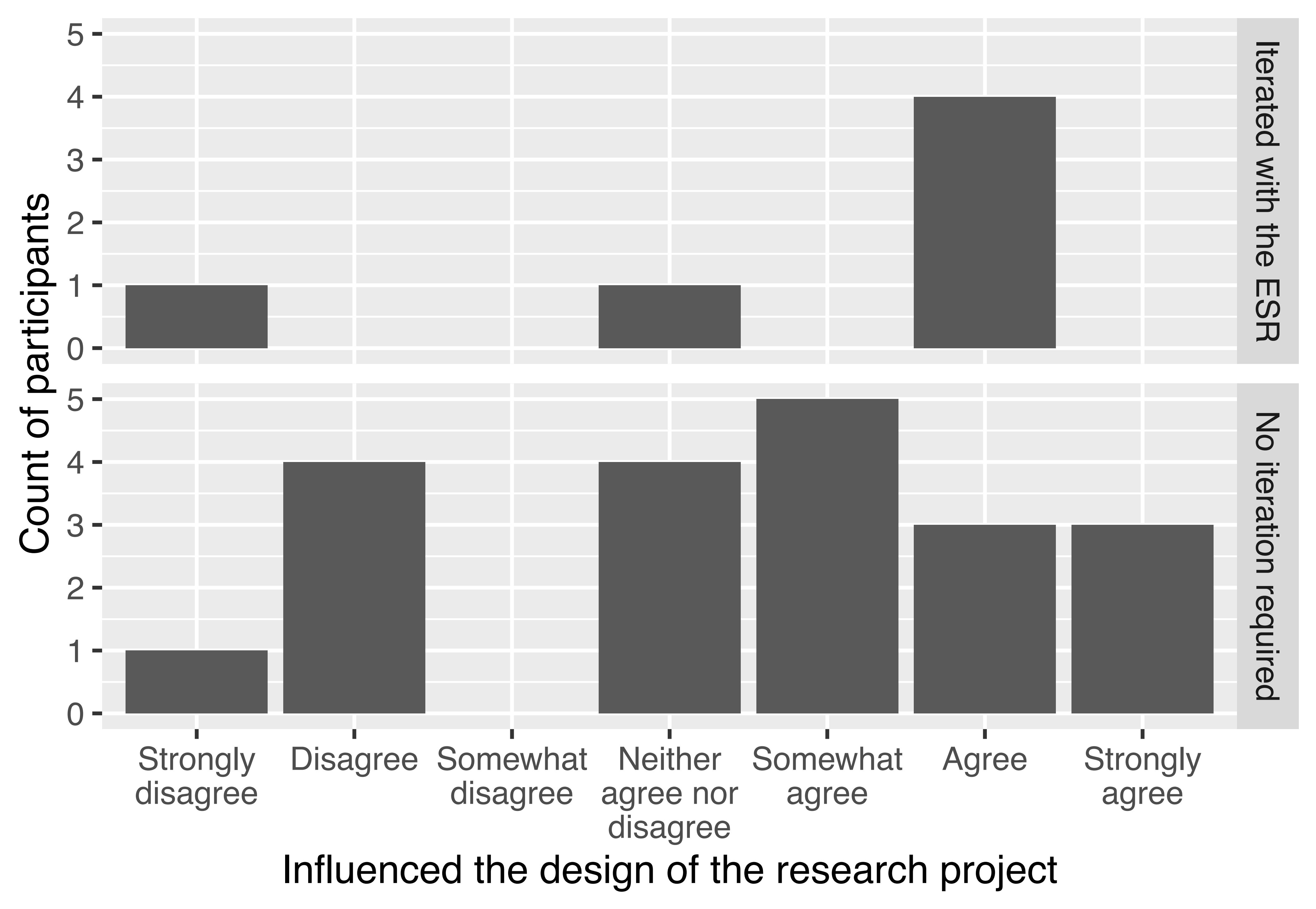
Feedback analyzed from consented **ESR panelists**

- 14/15 consented

Every survey respondent was willing to engage in the ESR process again

67% of those who iterated with the ESR, and 58% of all researchers, felt that the process had **influenced the design of their research**





THE MAIN BENEFIT: SCAFFOLDING

Researchers felt that the ESR served as a **forcing function and commitment mechanism** for thinking about ethics and societal impact

> "The ESR 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."
> — PI, Social Science

THE MAIN BENEFIT: SCAFFOLDING

Eight of thirteen interviewees said that the ESR raised **new ethical and social issues** for them to think about

> “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 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.
> — PI, Engineering

MAIN DRAWBACK: NOT ENOUGH SCAFFOLDING

Most consistent feedback: don’t just help broaden social and ethical lenses, also **provide scaffolding to make appropriate considerations**

The ESR statement was kept brief to keep work minimal, but **participants wanted more detail and specificity**

"[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 termsof 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.** - PI, Medicine

nature machine intelligence

The perils of machine learning in designing new chemicals and materials

Sadasivan Shankar & Richard N. Zare

Recently, our university's Ethics and Society Review panel reached out regarding one of our proposed projects, which involved the use of machine learning to predict the toxicity of chemicals and materials. The panel raised important questions about the ethics and societal consequences of our research. On the one hand, once perfected, this power could be used to scan for unwanted toxic materials – for example, in all the chemicals that are used in fracking fluids to extract oil. On the other hand, it could also be used by malicious actors to search for new ways to poison the ground or water. Specifically, the panel told us we should think about ways to control the distribution of the software, the model, and its output to minimize potential misuse.

After discussions with the panel, we sought the advice of other experts on how to overcome

Recommendation 6-2: The NAIRR should establish an ethics review process to vet all resources included in the system and the research performed within.

External ethics reviewers (as generally described in Chapter 3 of this report) should be leveraged for this purpose. While the majority of data in the NAIRR is not expected to have ethical concerns, the NAIRR management entity should establish and implement acceptance criteria and recommended best practices for all resources joining the NAIRR to ensure that they are vetted from privacy, civil rights, civil liberties, and inclusivity perspectives. This acceptance criteria should be more stringent for resources that are likely to be used in contexts that raise heightened concerns about privacy, civil rights, and civil liberties. In the case of third-party data sets made available via the NAIRR, this vetting process would need to be developed and could include establishing certification standards and/or providing trusted and validated reference data sets for testing (i.e., as an audit system). Only after appropriate vetting may these data sets be included in the NAIRR. In addition, the inclusion of higher risk data sets that have been modified with embedded privacy protections must be reviewed by potentially affected communities, because of the possible impact on those communities.

"**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." - PI, Social Science

COMMON THEMES IN ESR FEEDBACK

Harms to subgroups (31% of grants): which groups may be negatively impacted if this is widely adopted?

Diverse design (23%): are relevant stakeholders included?

Dual use (23%): how might this be reappropriated by motivated actors?

Representativeness (17%): who is in the data?

Panelists raised new issues for 80% of proposals that iterated with the ESR

OPEN QUESTIONS AND NEXT STEPS

How do we scale this process up to hundreds of proposals per year, and to other institutions? Can we do this while maintaining a coaching lens rather than a compliance lens?

How can we measure the impact of the ESR?

ESR: Ethics and Society Review of AI Research

Thanks to...

- Supporters: NSF, Stanford HAI
- Stanford Center for Advanced Study in the Behavioral Sciences
- Stanford Ethics, Science, and Technology Hub
- Questions