

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*, Paul G. Allen School of Computer Science & Engineering, University of Washington
 *Allen Institute for Artificial Intelligence
<https://rowanzellers.com/grover>

Abstract

Recent progress in natural language generation has raised dual-use concerns. While applications like summarization and translation are positive, the underlying technology also might enable adversaries to generate *neural fake news*: targeted propaganda that closely mimics the style of real news. Modern computer security relies on careful *threat modeling*: identifying potential threats and vulnerabilities from an adversary's point of view, and exploring defenses against these threats. Likewise, developing robust defenses against Grover. Grover can generate the rest of the article; humans find these generations to be more trustworthy than human-written disinformation. Developing robust verification techniques against generators like Grover is critical. We find that best current discriminators can classify neural fake news from real human-written news with 73% accuracy, assuming access to a moderate level of training data. Counterintuitively, the best defense against Grover turns out to be Grover itself, with 92% accuracy, demonstrating the importance of exposing strong generators. We investigate these results further, showing that artifacts of bias – and sampling strategies that alleviate its effects – both leave artifacts that similar discriminators can pick up on. We conclude by discussing ethical issues regarding the technology, and plan to release Grover publicly, helping pave the way for better detection of neural fake news.

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini
 MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru
 Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks. IJBA and Adience.

who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg^{a,1}, Londa Schiebinger^b, Dan Jurafsky^{c,d}, and James Zou^{a,1}

^aDepartment of Electrical Engineering, Stanford University, Stanford, CA 94305; ^bDepartment of History, Stanford University, Stanford, CA 94305; ^cDepartment of Linguistics, Stanford University, Stanford, CA 94305; ^dDepartment of Computer Science, Stanford University, Stanford, CA 94305; ^eDepartment of Biomedical Data Science, Stanford University, Stanford, CA 94305; and ^fChan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

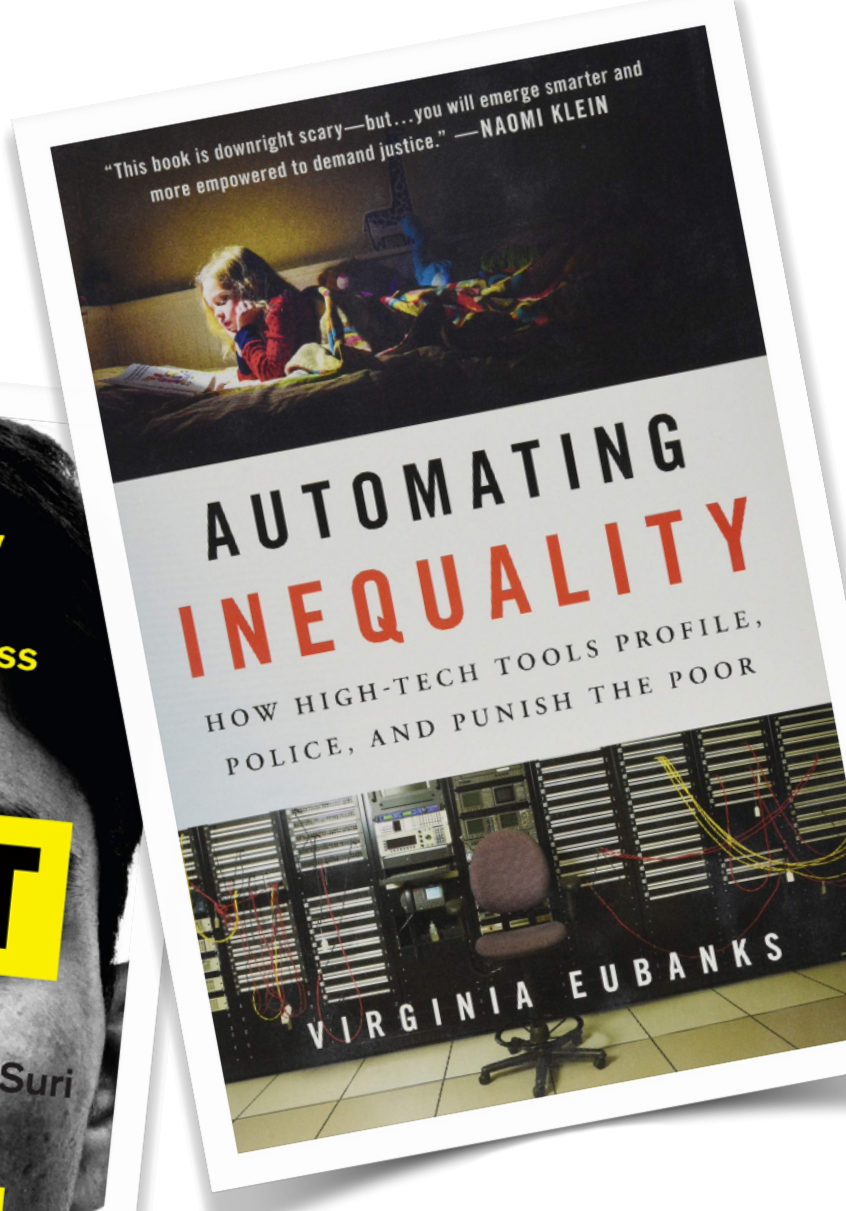
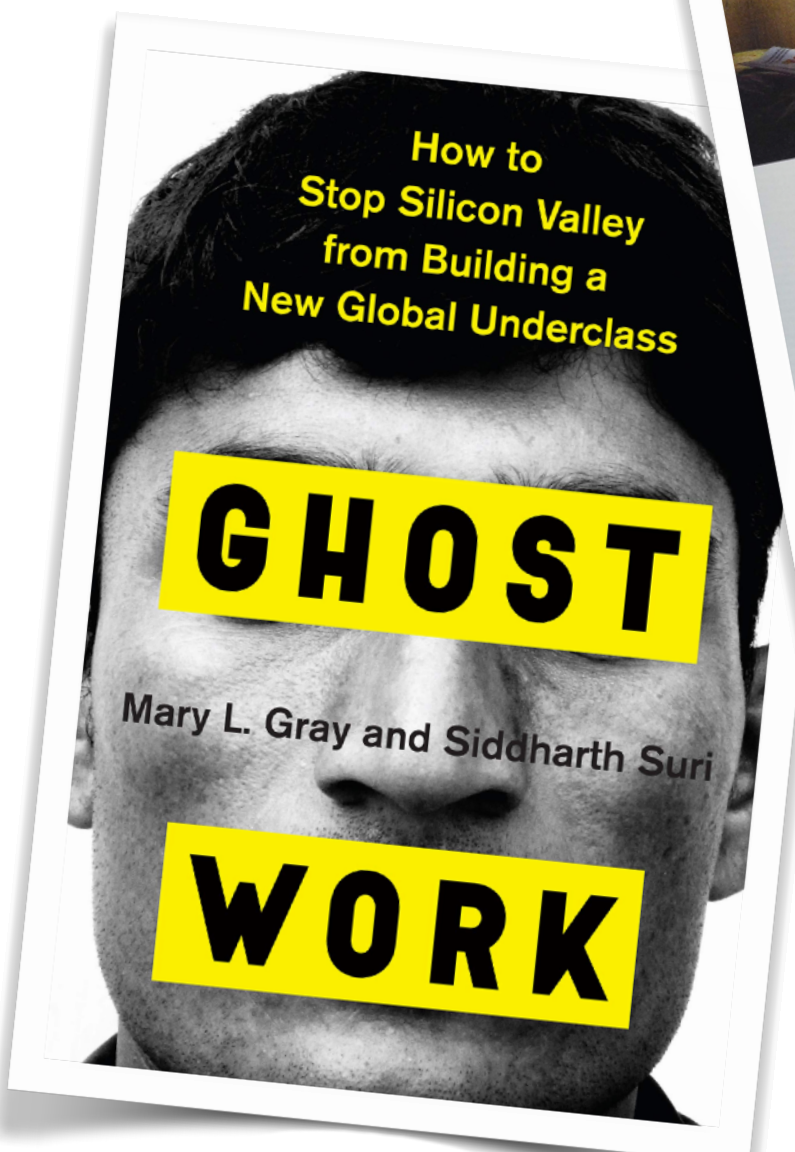
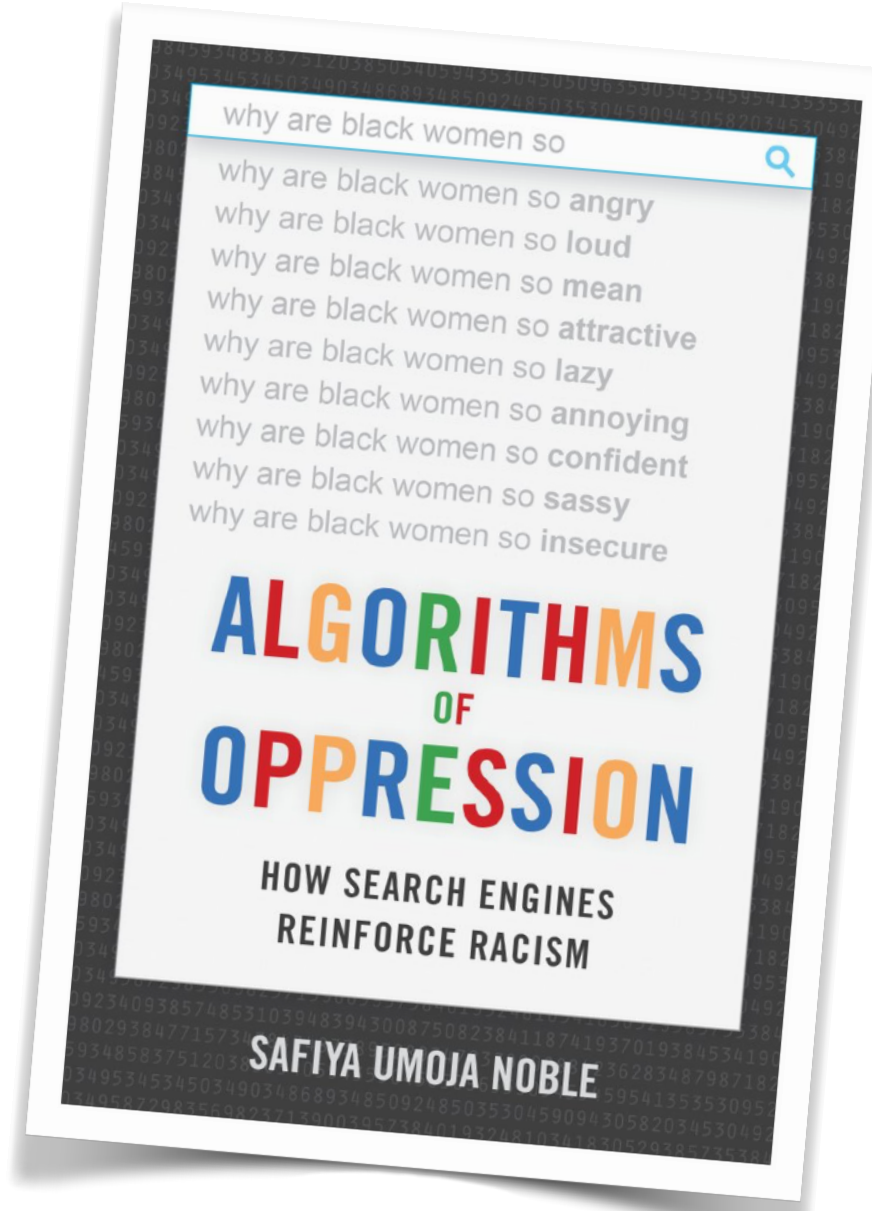
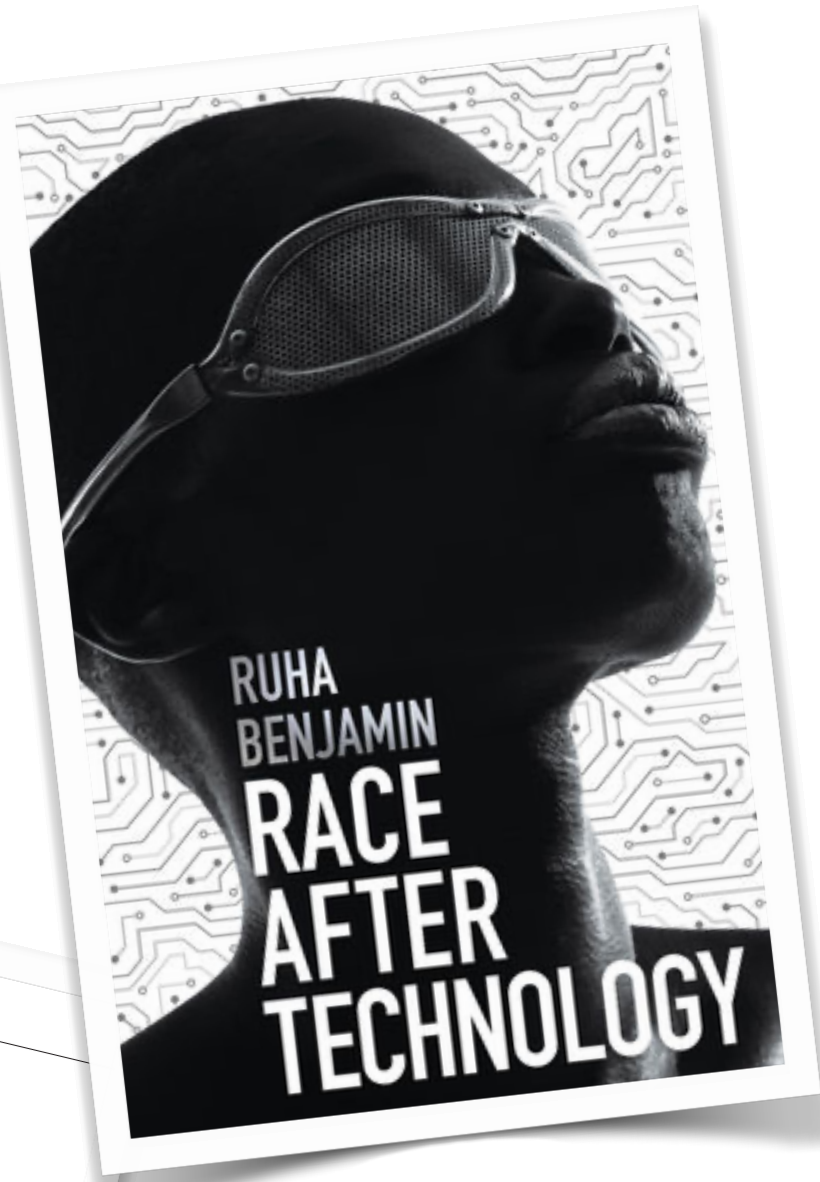
in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20). In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly

Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management

Min Kyung Lee

Abstract

Algorithms increasingly make managerial decisions that people used to make. Perceptions of algorithms' actual performance, can significantly influence their adoption, yet we do not fully perceive decisions made by algorithms as compared with decisions made by humans. To explore algorithmic management, we conducted an online experiment using four managerial decisions that require algorithmic or human skills. We manipulated the decision-maker (algorithmic or human), and measured emotional response. With the mechanical tasks, algorithmic and human-made decisions were fair and trustworthy and evoked similar emotions; however, human managers' fairness and trustworthiness were higher than algorithmic decisions. Human decisions evoked some positive emotion due to the perceived efficiency and objectivity. Human decisions evoked a more mixed response – algorithms were seen as more efficient but less trustworthy.



RESEARCH

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2*}, Brian Powers³, Christine Vogel⁴, Seshil Mullainathan^{5*,†}

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for illness, seemingly effective proxies for ground truth cost are biased in many contexts.

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

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets*. In the electronics industry, a component, no matter how simple or complex, is accompanied by a datasheet that describes its operation.

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

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

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
<p><i>Representativeness</i> Insufficient or unequal representation of data, participants, or intended user population</p> <p>Example: data collection process for a wellbeing sensing algorithm would undersample low-income populations</p>	<p>Algorithm training data and evaluation should include communities likely to be impacted by the algorithm</p>	<p>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</p>

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

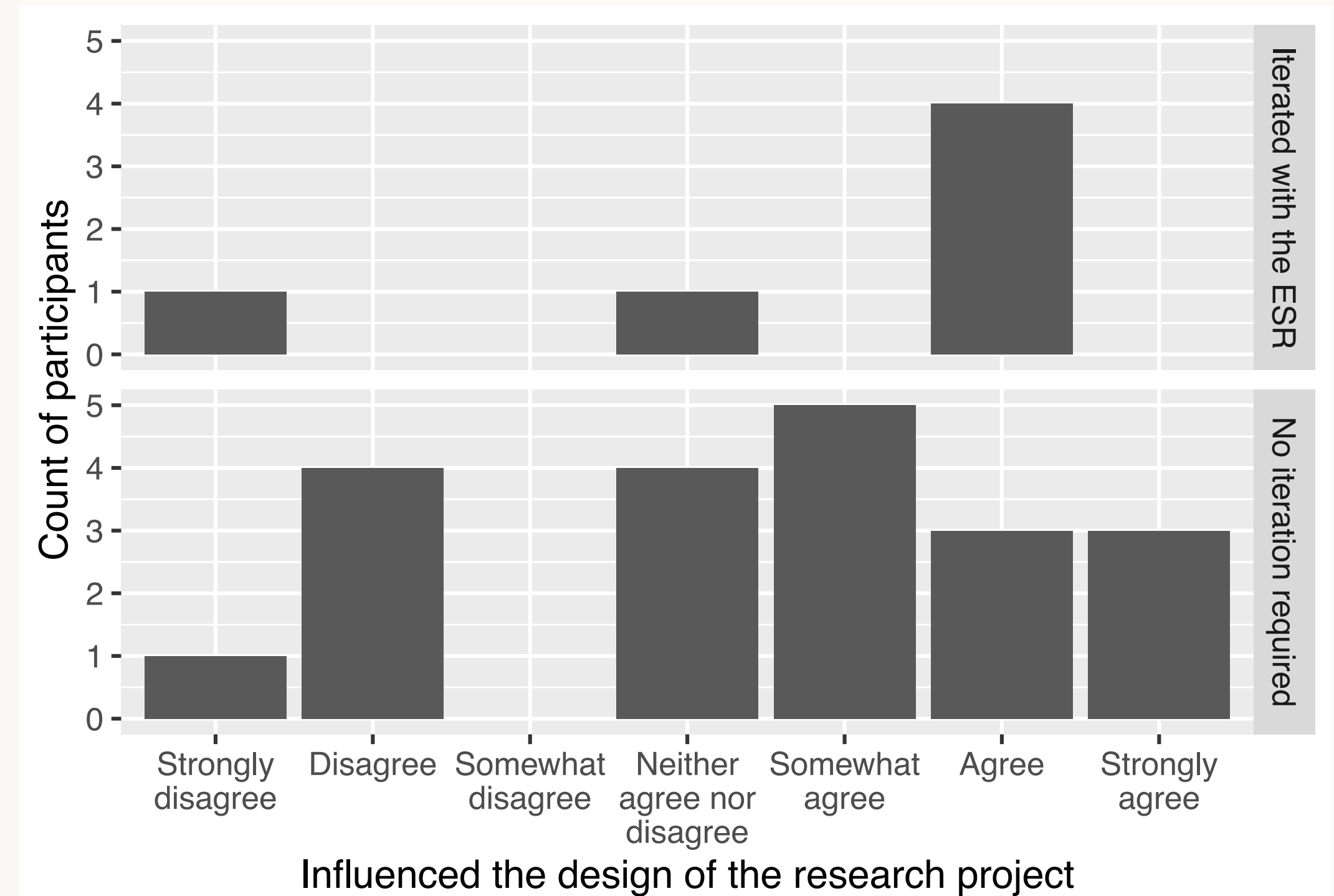
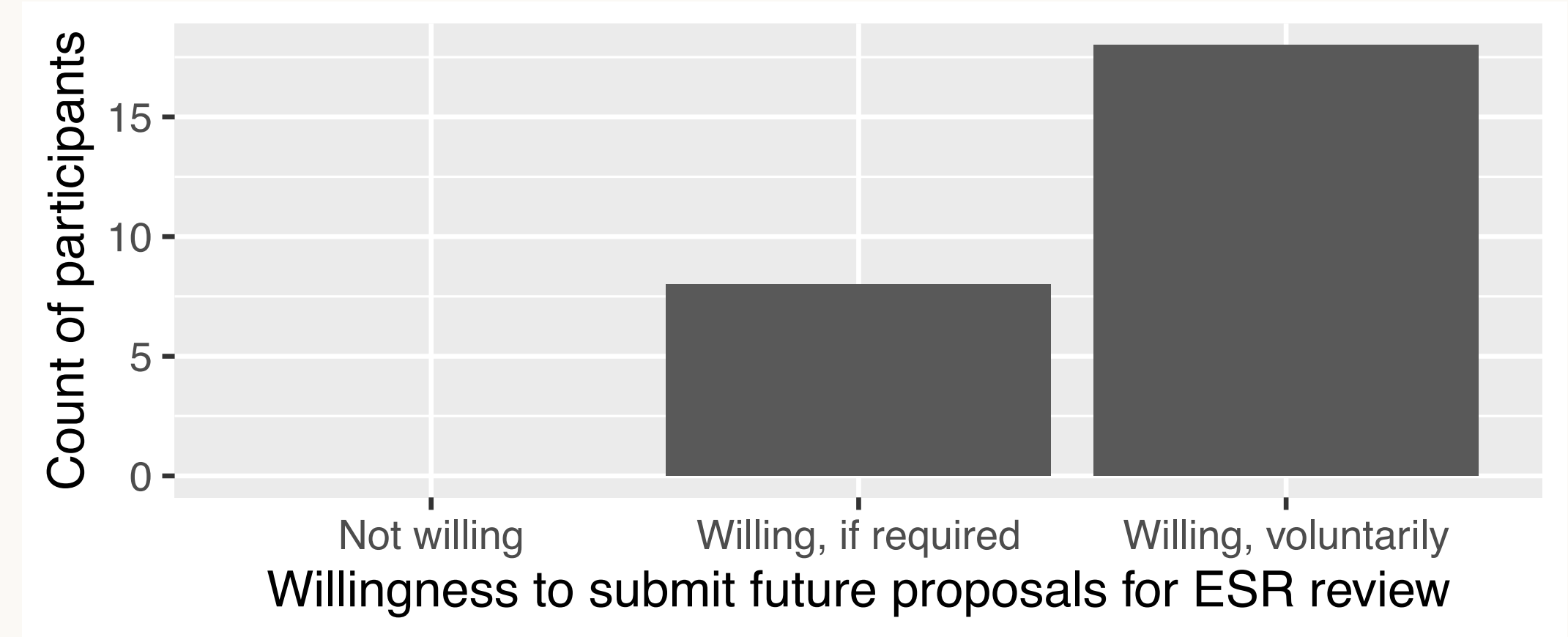
13 projects: Focuses in engineering (7), social science (4), earth science (2), and medicine (2)

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 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.** - PI, Medicine

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