A. STEVIE BERGMAN

Sociotechnical Research & Tech Policy

✉ bergman.as@gmail.com   🌐 astieviebergman.com

📍 New York City   in /astieviebergman

SUMMARY

Public interest scientist, technologist, and sociotechnical expert with years of experience in the tech industry and a strong commitment to human rights and equity. Areas of expertise include AI ethics, governance and tech policy, democratic and participatory AI, implications of generative AI, algorithmic fairness, mis/disinformation, sociotechnical AI evaluation, representation, multilingual AI, and data annotation, labor, and curation practices. Experience in diverse interdisciplinary areas such as legal work and international aid, physics research, and science & tech communication.

CURRENT POSITION

Senior Research Scientist   4/2022 – Present
Ethics Research Team, Google DeepMind   steviebergman@google.com

Rigorous sociotechnical research employing mixed methods on complex, open-ended problems across a host of interdisciplinary areas including: international **technology policy** and ethical AI development, **democratic and participatory** AI, sociotechnical **AI evaluation**, representation and fairness in AI, the impact of emerging technologies on non-Western, at-risk or in-conflict communities, **societal impacts of generative AI** (eg., **mis- and disinformation**), and the implications of design choices (eg., **data curation** and AI annotation with **fair labor practices**).

Role has involved
- Leading teams of colleagues (research scientists, engineers, etc) on complex, multi-stream research projects with several dependent deliverables and high-level buy-in,
- Expert consultation to legal and policy teams,
- Internal partnership with product and research teams across Google,
- External collaborations and engagements, such as close collaboration with the National Institute of Standards and Technology, providing feedback on the AI Risk Management Framework (2022-24), fireside chats, dialogues, workshops, and panels on participation in and governance of AI systems, represenatation in AI, as well as the many risks of disinformation from generative AI. See **Public Engagements**.

PRIOR ROLES

1/2020 – 4/2022   **Research Scientist, AI Fairness, Responsible AI**   **Meta**
Practical application of AI ethics principles to Meta products with a keen focus on countries and communities outside the West, and at-risk for offline harm. Flagship research focused on Arabic dialects, hateful and violent speech, annotation/labor practices (see: *Towards Responsible Natural Language Annotation for Varieties of Arabic*), and the dynamics technology design choices foment in the region. Additional research on African languages and Portuguese. Further work in this role included:
- Development of novel, responsible research methodologies to address open-ended questions and recommend high-impact mitigation strategies.
- Utilization of interdisciplinary research methods including qualitative & quantitative data analysis techniques, systematic root cause analysis, robust, ethical, and inclusive data collection & representative sampling, and surveys.
- Production of many internal research reports on these investigations & their impacts.
- Technical subject-matter expert on the Responsible AI Education team, producing and delivering clear, high-impact education projects, driving adoption across the company.
- Creating and co-leading weekly AI Ethics Foundational Reading & Discussion group.

Spring & Fall 2021   **Lecturer, Ethics of AI Graduate Seminar**   **Princeton University**
Seminar speaker on the practical application of AI ethics with a focus on methods for accountability & the complications involved with measurements for benchmarking. After taking this seminar in 2019, was invited back as a lead to teach on the topic.

Spring 2019   **Creator, Producer, and Host of AI & Human Rights Podcast Miniseries**   **Princeton University**
Educational podcast commissioned by Princeton's Center for Information & Technology Policy, on the intersection of digital technologies and human rights, framed by a CITP/UN conference in April 2019. Required research into the ethical implications of emerging technologies and a grasp of technology policy and its implications for society. Available on Apple, Spotify, & Soundcloud.

7/2015 – 10/2019   **Creator, Co-Host, Co-Producer of These Vibes Are Too Cosmic**   **WPRB Princeton**
Weekly, prime time science & music radio show featuring science news, local events, and live, longform interviews with experts from across the sciences. Full list of 100+ interviews available for streaming on thesevibesaretoocosmic.com.

| 10/2013 – 7/2014 | **Fulbright Fellow, Indonesia** | **Institut Teknologi Bandung** |
| | Independent research in theoretical high energy physics (complex scalar field dynamics with BPS domain walls), with university collaborator. | |

| Summer 2013 | **Research Assistant, Large Hadron Collider at CERN** | **Columbia University** |
| | Testing & commissioning the inner B-Layer of the ATLAS Pixel Detector in Geneva, Switzerland. | |

2/2011 – 4/2013	**US Peace Corps, Science Education Volunteer**	**Uganda**
	Mathematics, physics, and computer teacher at Lacor Secondary School and Warr Girls.	
	Creator & Co-director of GirlTech Uganda and Secondary Education PCV trainer.	

| 8/2009 – 12/2010 | **Vaccine Litigation Paralegal** | **US Department of Justice** |
| | Assisted attorneys in understanding the science involved in their work, as well as research, documentation, and argument for cases at the US Court of Federal Claims. | |

Other Projects

2020	Flatbush Mutual Aid, New York City, NY
2016 – 2019	Prison Teaching Initiative, Weekly Tutor, Garden State Prison, NJ
2018 – 2019	Physics Dept. Climate and Inclusion Committee, Princeton University

PUBLICATIONS IN TECH & SOCIETY

arXiv 2024
Submitted for publication

STAR: SocioTechnical Approach to Red Teaming Language Models
Laura Weidinger, John Mellor, Bernat Guillén Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, **Stevie Bergman**, Mikel Rodriguez, Verena Rieser, William Isaac
arxiv.org/pdf/2406.11757

Nature Scientific Reports 2024

STELA: A Community Centred Approach to Norm Elicitation for Agent Alignment
Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, William Isaac
nature.com/articles/s41598-024-56648-4

Released Apr 2024
Future submission for publication

Ethics of Advanced Assistants: Access and Opportunity (Chapter)
A. Stevie Bergman, Renee Shelby, Iason Gabriel
Full paper: the-ethics-of-advanced-ai-assistants-2024-i.pdf
Blog post: blog/the-ethics-of-advanced-ai-assistants

CHI 2024
Accepted for publication

The Illusion of Artificial Inclusion
William Agnew, **A. Stevie Bergman**, Jennifer Chien, Mark Dìaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, Kevin R. McKee
arxiv.org/pdf/2401.08572.pdf

arXiv 2023
Future submission for publication

Sociotechnical Safety Evaluation of Generative AI Systems
Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, **Stevie Bergman**, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser and William Isaac
Full paper: arxiv.org/pdf/2310.11986.pdf
Blog post: blog/evaluating-social-and-ethical-risks-from-generative-ai

ACM FAccT 2023

Representation in AI Evaluation
A. Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, William Isaac
dl.acm.org/doi/abs/10.1145/3593013.3594019

ACL 2022

Towards Responsible Natural Language Annotation for Varieties of Arabic
A. Stevie Bergman & Mona Diab
aclanthology.org/2022.findings-acl.31

ACM FAccT 2022

Adaptive Sampling Strategies to Construct Equitable Training Datasets
William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, **Stevie Bergman**, Sharad Goel
dl.acm.org/doi/abs/10.1145/3531146.3533203

SIGDIAL 2022

Guiding the Release of Safer E2E Conversational AI through Value Sensitive Design
A. Stevie Bergman, Gavin Abercrombie, Shannon Spruit, Dirk Hovy, Emily Dinan, Y-Lan Boureau, Verena Rieser
iris.unibocconi.it/bitstream/11565/4053244/1/2022.sigdial-1.4.pdf

ACL 2022

SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems
Emily Dinan, Gavin Abercrombie, **A. Stevie Bergman**, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, Verena Rieser
iris.unibocconi.it/bitstream/11565/4053224/1/2022.acl-long.284.pdf

arXiv 2021

Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems
Chloè Bakalar, Renata Barreto, **Stevie Bergman**, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, Jiejing Zhao
arxiv.org/pdf/2103.06172v2.pdf

Full list of journal articles available on Google Scholar, including publications in physics.

SELECT PUBLIC ENGAGEMENTS

May 2024 Keynote Speaker	**Third Workshop on Safety for E2E Conversational AI**	**LREC-COLING 2024 Conference**

Workshop focusing on the contextual and culturally dependent risks with end-to-end open-domain dialogue agents, such as not being able to respond reliably in safety critical situations. Keynote will be on the topic of culturally situated and sociotechnical evaluations of these models, as well as effective representation and community inclusion.

April 2024
Workshop

Operationalizing the NIST AI Risk Management Framework: A Diversity of Human Factors
                                          NIST & Northeastern University Ethics Institute
Invitation-only workshop bringing together experts to build on NIST's AI Risk Management Framework to work towards developing operational, reproducible, and scalable evaluation tools and standards for mapping and assessing critical aspects of human involvement throughout the lifecycle of AI systems.

November 2023
Presentation

AI Governance, Evaluations, and Representation   **Center for Democracy & Technology**
Invited presentation to CDT non-resident fellows on Google DeepMind policy planning, and sociotechnical research with policy implications.

October 2023
Fireside Chat

"Risk Management (or Measurement), for whom, why and how?"   **NIST & CASMI**
Fireside chat with Zachary Lipton, facilitated by Nicol Turner Lee of the Brookings Institution, at the NIST Workshop on Operationalizing the Measure Function of the AI Risk Management Framework. Organized by Northwestern Center for Advancing Safety of Machine Intelligence (CASMI).

October 2023
Workshop

Operationalizing the Measure Function of the NIST AI RMF   **NIST & CASMI**
Two day, invite-only workshop to discuss the complexities of putting the sociotechnical principles outlined in the AI Risk Management Framework into practice, organized by Northwestern Center for Advancing Safety of Machine Intelligence (CASMI).

July 2023
Panel

Potential Benefits and Risks of Generative AI   **Cambridge University, Disinformation Summit**
Panelist alongside representatives from McKinsey and Meta, fielding audience questions on the growing risks (and potential benefits) of generative AI – particularly mis/disinformation risks and threats to democracy. Full recording can be found here.

2023 (ongoing)
Dialogue

Public Technology Leadership Collaborative (PTLC)   **Data & Society**
A collaborative group of civil servants, industry technologists, and academics, exchange insights for responsible development and governance of AI technologies.

June 2023
Dialogue

Discussions between AI ethics and AI safety researchers   **ACM FAccT 2023**
Two part dialogue alongside the 2023 FAccT conference, between sociotechnical AI ethics researchers (who tend to engage with current risks of AI) and AI safety researchers (who tend to be far-long-term engaged).

June 2023
Presentation

Representation in AI Evaluations   **ACM FAccT 2023**
Presentation on the paper *Representation in AI Evaluation* at the Fairness Accountability and Transparency conference in Chicago, 2023.

April 2023
Lecture

Experience AI lesson on Data Bias   **Raspberry Pi & DeepMind**
Conversational lecture on bias in data for the Experience AI lessons for schoolchildren, in collaboration with Raspberry Pi. (Not yet public.)

April 2023
Conference &
Workshop

Envisioning equitable representation in AI   **Cambridge, Many Worlds of AI**
At this conference, given extended time to present an engaging workshop walking participants through the complexities and pitfalls in thinking through effective representation in AI design, and the consequences of different design choices. This workshop presented and extended FAccT 2023 publication on *Representation in AI Evaluation*.

October 2022
Workshop &
Presentation

Representation in AI Annotation   **REAL ML**
Invite-only work-in-progress workshop at which I presented research on considerations, complexities, and implications of representation in AI data annotation.

October 2022
Panel

Building the NIST AI Risk Management Framework (Workshop #3)   **NIST**
Panelist discussing participatory & sociotechnical methods for integration into the (at the time) upcoming AI Risk Management Framework from the National Institute of Standards and Technology. Recording can be found here.

July 2022
Dialogue

Dialogue on an Institution for Assessing Digital Harms   **New America Foundation**
Conversation across organizations and disciplines to explore the prospects for a trusted, neutral global institution that could develop agreed definitions and measurements of digital harms.

October 2022 Workshop	**Assessing Social Impacts of General Purpose AI Systems** Workshop bringing together academic and industry experts to brainstorm the requirements for effective sociotechnical evaluations of a generative AI systems.	**Hugging Face**
2019 (ongoing) Dialogue	**TechSoc Discussion Group** Monthly reading and discussion group on technology and society topics, organized by the Center for Information Technology Policy.	**Princeton CITP**
2020 (ongoing) Dialogue	**Privacy Research Group** Weekly discussion group on information law and technology policy topics – typically works in progress from group members – organized by New York University's Information Law Institute.	**NYU ILI**

EDUCATION

10/2019	**Doctorate in Physics** Observational cosmology & experimental physics. NSF Graduate Research Fellowship (2015-18), Joseph Henry Merit Prize (2014) **Thesis description:** Measuring the statistical polarization in the cosmic microwave background radiation from the early Universe among instrumental, environmental, and astrophysical noise, via constructing the most sensitive radio telescopes ever built, the Spider instrument, & flying it on a balloon above Antarctica. **Work included:** Close teamwork on an experiment with many potential single-point failures, data analysis and experimental research methods.	**Princeton University**
5/2009	**BA Physics, Minor in Astrophysics** cum laude, Highest Honors, Phi Beta Kappa, Sigma Xi, Waterman Prize for Outstanding Senior in Physics Year abroad – Oxford University	**Smith College**

LANGUAGES

Python, SQL, C/C++, some French, some Acholi and Alur, learning Levantine Arabic