THOMAS GILBERT
130 W 82ND St, Apt 2R
New York, NY 10024
Mobile: 510-388-4841
Email: thomaskrendlgilbert@gmail.com
LinkedIn: https://www.linkedin.com/in/thomas-krendl-gilbert-38425b31/

WORK EXPERIENCE

Reward Reports - New York, NY, USA

Project Lead - 02/2022 to Present - Hours per week: 20

Reward Reports are a framework for documenting the capabilities and societal risks of increasingly advanced AI systems. This framework makes it possible to track the dynamic and holistic behaviors of AI systems over time. With funding from the Mozilla Technology Fund, the Reward Reports project combines perspectives from law, computer science, social science, and philosophy to construct tools for deliberation on the optimization criteria of AI systems.

DUTIES AND RESPONSIBILITIES: Conducting user-centered needs analysis for policymakers tasked with oversight of AI technologies enabled by reinforcement learning (RL). Leading product development of interfaces for stakeholders (designers, vendors, clients, users) of RL-enabled systems. Forming partnerships with cities nationwide to integrate the Reward Reports workflow within relevant digital offices.

USER-CENTERED NEEDS ANALYSIS:

- Led an interview study in summer 2023 to solicit insights from policymakers and AI designers on the requisite components of Reward Reports. Presented findings at leading AI + Society research venues (FAccT, AIES).
- Wrote up findings from user study in an internal research memo currently prepared for publication in 2024.
- Hired a UX designer and backend API developer to build new technical infrastructure for Reward Reports in response to user study findings.

PRODUCT DEVELOPMENT:

- Partnered with the Mozilla Technology Fund ($50,000 award) to develop Auditing Tools for AI Systems (Jan. 2023-Present).
- Partnered with Cornell University ($10,000 award) to develop a GitHub Repository for Reward Reports (May 2022-Present).
- Released a Minimum Viable Product (MVP) of Reward Reports in Aug. 2023, available here: https://rewardreports.github.io/reward-reports/builder/index.html

CITY PARTNERSHIPS:

- Coordinating policy feedback on Reward Reports from the digital offices of Boston, New York City, Portland, and San Jose. Preparing pilot use of Reward Reports with these cities to begin Feb. 2024.

TECHNICAL SKILLS:

Drafting research outputs using Overleaf and LaTeX to generate research outputs. Leading the work of a technical team that uses Javascript, Python, and RL frameworks to build Reward Reports interface components.

SELECTED WORK:

- “Reward Reports for Reinforcement Learning” (with Nathan Lambert et al). *Proceedings of the AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*. 2023.
- “Dynamic Documentation for AI Systems” (with Soham Mehta and Anderson Rogers). *Proceedings of the Designing Technology and Policy Simultaneously workshop, CHI*. 2023.
- “Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems” (with Sarah Dean, Tom Zick, and Nathan Lambert), *Center for Long-Term Cybersecurity Whitepaper Series*, February 2022.
- “Axes for Sociotechnical Inquiry in AI Research” (with Sarah Dean et al.), IEEE Transactions on Technology and Society *2* (2), 62-70. 2021.
- “AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks” (with McKane Andrus et al.), IEEE International Symposium on Technology and Society. 2020.
- “A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics” (with Roel Dobbe et al.). *Proceedings of the FATML Workshop, ICML*. 2018.

The New York Academy of Sciences - New York, NY, USA

Consultant - 09/2023 to Present - Hours per week: 20

The mission of the New York Academy of Sciences (NYAS) is to drive innovative solutions to society’s challenges by advancing scientific research, education and policy. Among the oldest scientific organizations in the United States, the Academy is an important and widely-respected contributor to the international scientific community. To drive scientific progress, the Academy hosts over 150 conferences & symposia annually, connecting experts across sectors, disciplines, and national boundaries. The Academy further brings individuals and organizations together to drive real-world solutions to global challenges.

DUTIES AND RESPONSIBILITIES: Advising the executive leadership of NYAS on the launch of new AI + society research initiatives. Defining opportunities for user-centered research in the context of the rapid advancement of AI technologies (e.g., large language models, reinforcement learning from human feedback). Leading the development of NYAS policy to critically adjudicate these tensions through interdisciplinary convenings, public outreach, and re-articulations of democratic values in AI development.

LEADERSHIP ADVISING:

- Spearheading NYAS grants and fellowship applications (>$200k) to support and sustain the AI + Society initiative beyond the 2023-2024 cycle. Coordinating applications on behalf of executive leadership and relevant Board members.

- Convening regular meetings with executive leadership and the Program Manager to redefine NYAS's standing commitments to open science in the context of generative AI development. Providing input on relevant hiring decisions for future postdoctoral fellows and affiliated faculty to executive leadership.

USER-CENTERED RESEARCH INITIATIVES:

- Directing a weekly internal research seminar series attended by executive leadership, resident faculty, and postdoctoral fellows. Organizing guest speaker lists and hosting AI experts from the wider New York City research ecosystem (Cornell Tech, NYU, Columbia). Mentoring postdoctoral fellows on a weekly basis and consulting with them on select research topics, opportunities for collaboration, and journal submissions.
- Launching an external, public-facing seminar series in February 2024 on AI + Society and relaunching the annual NYAS Machine Learning Workshop Series in March 2024 after a five-year hiatus due to COVID-19. Research topics will include: automated vehicle fleets that optimize traffic flow; generative AI tools that transform student learning; recommender algorithms that affect how social media users consume and share content.

POLICY DEVELOPMENT:

- Gave public recorded testimony at New York City Hall in September 2023 on the NYAS policy towards new EdTech tools enabled by Generative AI.
- Proposing strategic paths for NYAS to develop an active stance as AI development continues while maintaining neutrality and openness to multiple points of view.
- Publishing op-eds on the White House Executive Order on Artificial Intelligence + UK AI Safety Summit for publication in leading news outlets (e.g. *Politico*, *Washington Post*, *New York Times*).
- Preparing whitepapers that propose frameworks for the evaluation and settlement of the inevitable value conflicts that emerge as AI systems are integrated into America life.

TECHNICAL SKILLS:

Providing strategic content for digital marketing, community outreach, and corporate innovation in the form of public op-eds, copywriting, style guides, plain language, comprehension/reading levels. Leading an interdisciplinary research team that uses Balsamiq, Google Drive, and Microsoft Office Suite (Word, Excel, Power Point).

SELECTED WORK:

- "Testimony from the New York Academy of Sciences on The Role of Artificial

Intelligence, Emerging Technology, and Computer Instruction in New York City Public Schools". Given at New York City Hall on September 20th, 2023.

- "AI and the EU Digital Markets Act: Addressing the Risks of Bigness in Generative AI" (with Ayse Gizem Yasar et al). *arXiv preprint arXiv:2308.02033* (2023).
- "Entangled Preferences: The History and Risks of Reinforcement Learning and Human Feedback" (with Nathan Lambert and Tom Zick). *arXiv preprint arXiv:2310.13595* (2023).
- "Open problems and fundamental limitations of reinforcement learning from human feedback" (with Stephen Casper et al). *arXiv preprint arXiv:2307.15217* (2023, under review at *TMLR*).

daios - New York, NY, USA

AI Ethics Lead - 09/2021 to Present - Hours per week: 8

The daios engine helps developers fine-tune ethical values into large language models based on user feedback, or personalized AI ethics. Our platform creates a feedback loop between users, data, and companies. Users can see how current training data influences AI behavior and give feedback. This feedback is used to build curated datasets to fine-tune models with values that are good for both the users and the companies that create them. The daios solution is value-agnostic in order to better reflect our complex reality and give people the means to choose their own morality.

DUTIES AND RESPONSIBILITIES: Advising the executive leadership of daios on the launch of new ethical AI technologies. Leading financial applications to support daios deep tech projects. Defining policy for daios in the AI ethics space.

LEADERSHIP ADVISING:

- Generating quarterly written assessments of the daios workflow for executive leadership.
- Assembling job descriptions and making recommendations to executive leadership for new positions.
- Hired a graphic designer for the daios website and whitepaper on behalf of executive leadership. Oversaw iterative design work and approved graphic content at distinct stages of development through five rounds of revision (mockup, wirehead, layout, graphics, complete copy) based on executive leadership goals.
- Facilitated and co-organized the in-person launch (50+ attendees) of the daios whitepaper in February 2023 at Hi-Note, NYC.

FINANCIAL AND POLICY DEVELOPMENT:

- Spearheaded and won the 2022 Notre Dame Tech Ethics Lab Award ($25,000) on behalf of daios. Inaugural cohort.

- Lead author of the daios whitepaper, distinguishing daios positions from other major approaches in the field of AI + tech ethics.
- Supplied data inputs and ethical feedback to finetune the behavior of the team's in-house large language model.

TECHNICAL SKILLS:

Provided strategic content with skills in copywriting, style guides, plain language, comprehension/reading levels. Worked on a cross-functional team that used Figma, Slack, Photoshop, Google Drive, and Microsoft Office Suite (Word, Excel, Power Point).

SELECTED WORK:

- "Beyond Bias and Compliance: Towards Individual Agency and Plurality of Ethics in AI" (with Megan Brozek and Andrew Brozek), 2023 (https://www.daios.tech/whitepaper).

Cornell Tech, Digital Life Initiative - New York, NY, USA

Postdoctoral Fellow - 09/2021 to 08/2023 - Hours per week: 40

The Digital Life Initiative explores societal perspectives surrounding the development and application of digital technology, focusing on ethics, policy, politics, and quality of life. Housed at Cornell Tech, the lab issues master's and doctoral degrees to scholars working at the intersection of AI, philosophy, and the social sciences.

DUTIES AND RESPONSIBILITIES: Led the development of data policy for societal-scale recommender systems. Designed solutions from the field of mechanism design for feedback-laden digital platforms. Developed data analytics strategies for randomized-controlled trials undertaken by Big Tech companies. Worked as a member of an interdisciplinary lab including AI designers, roboticists, legal theorists, and policy entrepreneurs to articulate new research directions for AI & Society.

POLICY FOR MECHANISM DESIGN AND DATA ANALYTICS:

- Developed a critique of content moderation for recommender systems.
- Presented relevant research findings to campus-wide weekly seminars in February 2022 and March 2023.
- Outlined strategy for public outreach on these topics, culminating in coauthored pieces in *The Atlantic* and *MIT Technology Review*.
- Preparing a public event at the Berkman Klein Center on "Public Health for AI-Enabled Recommender Systems" to be held in February 2024.
- Awarded $25,000 through the Simons Institute to hold a Summer Cluster on AI and Humanity in 2022 ($25,000). Hosted cluster attended by 50 people, with 15 sessions, on five technical topics (AI documentation, fairness, audits, reinforcement learning, mechanism design).

RESEARCH AGENDAS ORGANIZED:

- Convened campus-wide research seminars on topics in short- and long-term AI governance.

- Coalesced novel interdisciplinary research collaborations at the intersection of AI Ethics and AI Safety. Co-presented resulting research outputs at FAccT, the world's leading interdisciplinary venue for AI Ethics.
- Spearheaded and led the "Automated Vehicles for Social Good" research cluster. Led related simulation research and literature reviews conducted by four Cornell Tech master's students.
- Organized two workshops at leading AI conferences: Designing Technology and Policy Simultaneously (CHI 2023); Building Accountable and Transparent RL (RLDM 2022).

TECHNICAL SKILLS:
Generated research outputs with skills in copywriting, style guides, plain language, comprehension/reading levels. Led research collaborations by using appropriate Python libraries and packages for reinforcement learning. Worked on a cross-functional interdisciplinary lab that used Slack, pattern libraries, Google Drive, Microsoft Office Suite (Word, Excel, Power Point), and semantic HTML.

SELECTED WORK:

- "Optimization's Neglected Normative Commitments" (with Benjamin Laufer and Helen Nissenbaum). *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023.
- "Accountability Infrastructure: How to implement limits on platform optimization to protect population health" (with Nathaniel Lubin). *arXiv preprint arXiv:2306.07443*. 2023.
- "We've Been Thinking About the Internet All Wrong" (with N. Lubin), *The Atlantic*. June 21, 2023.
- "Social media is polluting society. Moderation alone won't fix the problem" (with N. Lubin), *MIT Technology Review*. August 9, 2022.
- "Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice" (with Qian Yang et al.). *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
- "Fleets on the streets: how number, affiliation and purpose of shared-lane automated vehicle convoys influence public perception and blame" (with Noah Zijie Qu, Wendy Ju, & Jamy Li). *Transportation research part F: traffic psychology and behaviour*. February 2023.
- "Sociotechnical Specification for the Broader Impacts of Autonomous Vehicles" (with Aaron Snoswell, Michael Dennis, Rowan McAllister, & Cathy Wu). *Proceedings of the Fresh Perspectives on the Future of Autonomous Driving workshop, ICRA*. 2022.

SELECTED ACADEMIC GRANTS AND HONORS

H2H8 Association Grant, UC Berkeley, 2021.

AI Security Initiative Summer Research Stipend, UC Berkeley, 2021.

Simons Institute Law and Society Fellowship, UC Berkeley, Fall 2020. Inaugural recipient.

Newcombe Fellowship, UC Berkeley, 2019-2020.

Best Poster Award, NeurIPS AI for Social Good Workshop, December 2019.

Center for Long-Term Cybersecurity Grant, UC Berkeley, 2019.

Research appointment with Prof. Stuart Russell, 2018-2019.

Social Science Matrix Research Award, 2018-2019.

Center for Long-Term Cybersecurity Grant, UC Berkeley, 2018.

Summer Intern at Center for Human-Compatible AI, UC Berkeley, 2018.

Research appointment with Prof. Ken Goldberg, 2016-2017.

SELECTED VOLUNTEER WORK AND PUBLIC SERVICE

Political Economy of Reinforcement Learning Systems (PERLS) – Berkeley, CA, USA

Founder and Conference Director - 09/2020 to Present - Hours per week: 5

PERLS is a cross-disciplinary group of researchers examining the near-term policy concerns of Reinforcement Learning (RL). PERLS explores both technical and institutional aspects of this problem, uncovering new research questions and methods of investigation. https://perls-group.github.io

DUTIES AND RESPONSIBILITIES: Awarded and served as the inaugural Law and Society Fellow at the Simons Institute in support of fall 2020 program on "Theory of Reinforcement Learning." Created and convened an international reading group on the social impacts of reinforcement learning systems across North America and Europe (130+ participants). Assembled the first ever syllabus on concepts and case studies in the political economy of reinforcement learning, based on regular group meetings with leading AI researchers and policy experts.

Graduates for Engaged and Extended Scholarship in Engineering (GEESE) – Berkeley, CA, USA

Research Co-Lead Dec. 2017–Present

GEESE is a collective of researchers from across the UC Berkeley ecosystem, interested in critically and constructively engaging the interfaces between technology and society. It now comprises a cross-institutional network of scholars and advocates working in industry, public policy, and academia. geesegraduates.org

DUTIES AND RESPONSIBILITIES: Executing a 2023 Mozilla Technology Fund Award: Auditing Tools for AI Systems ($50,000) to develop Reward Reports as Project Lead, comprised of GEESE researchers. Published a whitepaper on machine learning's societal risks with support from the AI Security Initiative at the Center for Long-Term Cybersecurity ($45,000).

Additional Selected Speaking and Presentations

"Hard Choices in Artificial Intelligence," Simons Institute for the Theory of Computing, UC Berkeley, July 13, 2022. Available on YouTube: https://www.youtube.com/watch?v=IouHJ6UxdUw&t=629s.

"Autonomous Vehicle Fleets as Public Infrastructure" (with Roel Dobbe), WeRobot, September 25, 2021.

"'Autonomous Vehicles' as Public Utility: Diagnosing Political Challenges Through the Lens of Normative Indeterminacy," Society for the Social Studies of Science (4S) conference, August 21, 2020.

"The Problem of Vagueness in Artificial Intelligence," Center for Human-Compatible AI 2020 Workshop, UC Berkeley, June 9, 2020.

"Bias in Datasets and Fairness in Machine Learning," Code for America Summit, May 30, 2019.

"Seeing Like an Algorithm: Machine Learning and the New Division of Apperceptive Labor," 4S annual conference, September 2, 2017.

Language Skills

Language: French
Spoken Level: Novice
Written Level: Novice
Reading Level: Intermediate

Language: German
Spoken Level: Novice
Written Level: Novice
Reading Level: Intermediate

Language: Danish
Spoken Level: Novice

Written Level: Novice
Reading Level: Intermediate

EDUCATION

Ph.D. University of California-Berkeley, Berkeley, CA Aug. 2021
Program in Machine Ethics and Epistemology
Fields: History and Theory of AI, Moral Cognition, Technology and Delegation
Dissertation: Modes of Deliberation in Machine Ethics

M.A. University of California-Berkeley, Berkeley, CA Sept. 2015
Sociology

M.Phil. University of Cambridge, Cambridge, UK June 2013
Political Thought and Intellectual History

Fulbright University of Copenhagen, Copenhagen, DK May 2012
Project: "Kierkegaard's Social Context in Historical and Contemporary Denmark"

B.A. Northwestern University, Evanston, IL June 2011
Sociology and Philosophy
Best Senior Thesis award

ADDITIONAL STUDY

Non-degree Danish Institute for Study Abroad, European Culture and History, 2010

REFERENCES

Nathaniel Lubin
Assembly Fellow
Berkman Klein Center
1557 Massachusetts Ave,
Cambridge, MA 02138
nlubin@cyber.harvard.edu

David Grewal
Professor
Berkeley Law School
225 Bancroft Way
University of California, Berkeley
david.grewal@berkeley.edu

Stuart Russell
Professor
EECS Department
387 Soda Hall,
University of California, Berkeley
russell@cs.berkeley.edu

Tom Zick
Berkman Klein Fellow
Berkman Klein Center
1557 Massachusetts Ave,
Cambridge, MA 02138
tzick@cyber.harvard.edu