

## Thomas Krendl Gilbert

Center for Human-Compatible AI  
University of California, Berkeley  
Berkeley, CA 94720

1920 Francisco Street, Apt. 103, Berkeley, CA 94709  
tg340@berkeley.edu  
(510) 388-4841

### EDUCATION

- Ph.D.** Program in Machine Ethics and Epistemology, expected August 2021  
University of California-Berkeley, Berkeley, California  
Fields: History and Theory of AI, Moral Cognition, Technology and Delegation  
Dissertation: Mapping the Political Economy of the Autonomous Vehicle Industry
- M.A.** Sociology, 2015  
University of California-Berkeley, Berkeley, California
- M.Phil.** Political Thought and Intellectual History, 2013  
University of Cambridge, Cambridge, UK  
Dissertation: "Kant's 'History of Pure Reason' and Construction of a Philosophical Persona in the *Critique of Pure Reason*"
- Fulbright Scholar** Søren Kierkegaard Research Centre, 2011-2012  
University of Copenhagen, Copenhagen, DK  
Project: "Kierkegaard's Social Context in Historical and Contemporary Denmark"
- B.A.** Sociology and Philosophy, 2011  
Northwestern University, Evanston, IL  
Senior Thesis: "The Social Construction of Søren Kierkegaard: A Modification of Randall Collins' Interaction Ritual Chains" (departmental honors)

### ADDITIONAL STUDY

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

### PUBLICATIONS

"Hard Choices in Artificial Intelligence" (with Roel Dobbe and Yonatan Mintz), *Artificial Intelligence* (under review)

"Subjectifying Objectivity: Delineating Tastes in Theoretical Quantum Gravity Research" (with Andrew Loveridge), *Social Studies of Science* (forthcoming).

“Epistemic Therapy for Bias in Automated Decision Making” (with Yonatan Mintz), *Proceedings of the 2019 AAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES '19)*: 61-67.

“Towards a Just Theory of Measurement: A Principled Social Measurement Assurance Program for Machine Learning” (with McKane Andrus), *Proceedings of AIES '19*: 445-451.

Review of Walter Lowrie, Kierkegaard, in *Kierkegaard Research: Sources, Reception and Resources: Volume 18*. Routledge 2016: 37-40.

Review of David Swenson, Something About Kierkegaard, in *Kierkegaard Research: Sources, Reception and Resources: Volume 18*. Routledge 2016: 245-248.

“Elias and the Sociology of Ideas: A Critique of Randall Collins’s Microsociology of Intellectual Change,” *Human Figurations* 5:1, March 2016.

“The Shipwreck of All Hopes: Liberalism and the Politics of the American Left” (with Andrew Loveridge), *Berkeley Journal of Sociology* 60, pp. 98-109.

“Why a Danish Golden Age? Structural Holes in Nineteenth-Century Copenhagen,” *Kierkegaard Studies Yearbook* 2013: 403-434.

“Heiberg’s Hegelianism: A Sociological Perspective,” *Kierkegaard Studies Yearbook* 2012: 201-234.

“Problems of Kierkegaard’s Poetics,” *Episteme* 22:5, 38-53.

### **WORKING PAPERS**

“AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks” (with McKane Andrus et al.). Submitted to the IEEE International Symposium on Technology and Society 2020.

“Technical Problems in Autonomous Vehicle Policy Research” (with Michael Dennis and Rowan McAllister). Submitted to the Navigating the Broader Impacts of AI Research Workshop, NeurIPS 2020.

“Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims” (with Miles Brundage et al.). Available on arXiv.

“A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics” (with Roel Dobbe et al.). Presented at the FATML Workshop at the 2018 International Conference on Machine Learning. Available on arXiv.

### **INVITED PRESENTATIONS**

“Hard Choices in Artificial Intelligence,” Simons Institute Law and Society Fellow talk, UC Berkeley, September 10, 2020. Available on YouTube: <https://www.youtube.com/watch?v=0n2NhKqpNCw>

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

“Søren Kierkegaard and the Limits of Randall Collins’s *The Sociology of Philosophies*,” Faculty of Sociology, University of Copenhagen, April 11, 2012.

“Framing the *Fragments*: Johannes Climacus and Rhetorical Subversion,” Kierkegaard Seminar at the University of Copenhagen, March 22, 2012.

### **CONFERENCE PRESENTATIONS**

“‘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.

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

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

“Kierkegaard as Political Theorist: Self-Formation as a Prelude to State Formation,” South Atlantic Modern Language Association annual meeting, November 5, 2016.

“A Comparative Social Morphology of Scientific Judgment in Theoretical Physics,” American Sociological Association annual meeting, August 21, 2016.

“The Historical Dynamics of Political Sublimation in Buddhism, Schopenhauer, and Wagner,” Religion and Irreligion in the History of Political Thought, University of Cambridge, May 25, 2015.

“Elias and the Sociology of Ideas: Kierkegaard’s *Either/Or* as the Psychogenesis of Existentialism,” Reinventing Norbert Elias: For an Open Sociology, Amsterdam, June 23, 2012.

“Is Existentialism a Civilizing Process? Elias, Kierkegaard, and the Sociology of Ideas,” Norbert Elias and Figurational Sociology: Prospects for the Future, Copenhagen, April 3, 2012.

“Problems of Kierkegaard’s Poetics,” OSU Philosophy Undergraduate Conference, May 13, 2011.

“Kierkegaard and the Sociology of Ideas: Pseudonyms as a Network of Intellectual Selfhood,” Chicago Area Undergraduate Research Symposium, April 2, 2011.

### **COURSES DESIGNED**

Spring 2020	<b>Foundations for Beneficial AI</b> (graduate level, co-taught with Prof. Stuart Russell)
Spring 2018	Demanding Reason (undergraduate composition class)

**COURSES TAUGHT**

Fall 2017	History and Theory of Reason
Spring 2017	Sense and Sensibility of Science
Spring 2016	Modernization and Development
Fall 2015	Sociological Methods
Fall 2014	Introduction to Sociology

**GRANTS AND HONORS**

**Simons Institute Law and Society Fellowship**, UC Berkeley, Fall 2020 (\$16,470). Inaugural recipient.

**Newcombe Fellowship**, UC Berkeley, 2019-2020 (\$25,000 plus fee remission).

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

Center for Long-Term Cybersecurity Grant, UC Berkeley, 2019 (\$37,000).

Research appointment with Prof. Stuart Russell, 2018-2019 (\$12,000).

Social Science Matrix Research Award, 2018-2019 (\$5,000).

Center for Long-Term Cybersecurity Grant, UC Berkeley, 2018 (\$15,000).

Summer Intern at Center for Human-Compatible AI, UC Berkeley, 2018 (\$6,000).

Research appointment with Prof. Ken Goldberg, 2016-2017 (\$8,000).

Doctoral position, Alpen-Adria-Universität Klagenfurt, Austrian Science Fund, 2016-2019 (declined).

Austrian Marshall Plan Foundation Grant, Summer 2016 (\$5,000).

Alumni Prize for Public Sociology, Berkeley Journal of Sociology Collective, 2015.

Dissertation Proposal Development Fellowship, Social Science Research Council, 2015 (\$2,000).

Small Research Grant Award, Department of Sociology, 2013 (\$2,000).

Research appointment with Prof. Jennifer Johnson-Hanks, 2013-2015 (\$12,000).

Research appointment with Prof. Marion Fourcade, 2014-2015 (\$8,000).

Research appointment with Prof. Raka Ray, 2014-2015 (\$4,000).

**Fulbright Research Award**, J. William Fulbright Foreign Scholarship Board, 2011-2012 (\$20,000).

Best Senior Thesis in Sociology, Northwestern University, 2011. Inaugural recipient.

Fletcher Family Academic Year Undergraduate Research Grant, 2011 (\$2,000). Inaugural recipient.

### **ACADEMIC SERVICE AND COMMUNITY INVOLVEMENT**

**Theory of Reinforcement Learning Program Fellow**, Simons Institute, Fall 2020

Authoring whitepaper on “Political Economy of Reinforcement Learning” (in process).

Projects Co-organizing: “Causal Inference,” “Political Economy of Reinforcement Learning.”

**Research Affiliate at Center for Human-Compatible AI**, UC Berkeley, 2017-Present

Authoring internal memos on social science research topics for technical interns.

Convening interdisciplinary seminars on political economy for AI theorists + Ph.D. students.

**Co-Founder of Graduates for Engaged Scholarship around Engineering (GEESE)**, 2017-Present

Co-authoring whitepaper on “Pedagogy Reforms for Engineering Ethics” (in process).

Designed survey of 90+ French engineering students on the future of AI and society.

Convened seminar on “Comparing the Politics of Computer Vision” at Berkeley Matrix.

Active Reviewer: *AI Ethics & Society*, *AI Magazine*, *American Journal of Sociology*

Active Member of Society for the Social Studies of Science (4S)

AI Vulnerabilities Workshop Participant, New York City, Fall 2019

Summer Institute on AI & Society Participant, Edmonton, Summer 2019

Philosophy and Physical Computing Workshop Participant, Virginia Tech, 2018

Graduate Student Working Group in Critical Realism Participant, Yale University, 2016-2017

Philosophy of the Social Sciences Graduate Student Summer Seminar Participant, 2015

Summer Fellow, Hong Kierkegaard Library at St. Olaf College (2011, 2013, 2019)

Graduate Student Editor, *Berkeley Journal of Sociology*, 2013-2016

**REFERENCES**

David Bates, chair  
Professor  
Department of Rhetoric  
7315 Dwinelle Hall,  
University of California, Berkeley  
dwbates@berkeley.edu

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