Jacy Reese Anthis

anthis@uchicago.edu

Research Interests: Social computing, human-computer interaction, machine learning, natural language processing, AI agents, AI safety, responsible AI, trustworthy AI

EXPERIENCE

2024–Current
Stanford University
Department of Computer Science and Institute for Human-Centered AI (HAI)
Visiting Scholar
Hosted by Michael Bernstein and Erik Brynjolfsson

2016–Current
Sentience Institute
Co-Founder and Director
501(c)(3) nonprofit research lab studying digital minds

June–Sept. 2025
Microsoft
Fairness, Accountability, Transparency, and Ethics (FATE)
Student Researcher
Modeling the predictability and human-likeness of AI errors

March–June 2025
Google
Technology, AI, Society, and Culture (TASC)
Student Researcher
Building synthetic data and evaluations of anthropomorphic LLM behavior

2023–2024
University of California, Berkeley
Department of EE&CS and Center for Human-Compatible AI (CHAI)
Visiting Scholar
Hosted by Stuart Russell

2015–2016
Animal Charity Evaluators
Researcher
501(c)(3) nonprofit organization that evaluates animal welfare programs

2014–2015
Harvard University
Research Assistant
Advised by Joshua Greene (moral cognition)

May–August 2014
GiveWell
Research Intern
501(c)(3) nonprofit organization that evaluates health and economic programs

2013–2014
University of Texas
Research Assistant
Advised by Russell Poldrack (cognitive neuroscience)

EDUCATION

Current
University of Chicago
Ph.D. in Sociology and Statistics (joint degree)

2021
University of Chicago
M.A. in Sociology

2015
University of Texas
B.S.A. in Neuroscience

CONFERENCE PAPERS

Notes: Conferences are the top peer-reviewed venues in human-computer interaction, machine learning, and natural language processing. † denotes that I served as principal investigator.

ACL 2025
The Impossibility of Fair LLMs [ACL, arXiv]
J. R. Anthis, K. Lum, M. Ekstrand, A. Feller, and C. Tan
Annual Meeting of the Association for Computational Linguistics

ACL 2025
Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation [ACL, arXiv]
K. Lum, J. R. Anthis, K. Robinson, C. Nagpal, and A. D'Amour
Annual Meeting of the Association for Computational Linguistics

ICML 2025
LLM Social Simulations Are a Promising Research Method [arXiv]
J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, E. Brynjolfsson, J. Evans, and M. Bernstein
International Conference on Machine Learning

FAccT 2025
What Do Large Language Models Say About Animals? Investigating Risks of Animal Harm in Generated Text [DOI, arXiv]
A. Kanepajs, A. Basu, S. Ghose, C. Li, A. Mehta, R. Mehta, S. D. Tucker-Davis, B. Fischer, and J. R. Anthis†
ACM Conference on Fairness, Accountability, and Transparency

CSCW 2025
The AI Double Standard: Humans Judge All AIs for the Actions of One [DOI, arXiv]
K. Manoli, J. V. T. Pauketat, and J. R. Anthis†
ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing

CHI 2025
Robots, Chatbots, Self-Driving Cars: Perceptions of Mind and Morality Across Artificial Intelligences [DOI, arXiv]
A. Ladak, M. Wilks, S. Loughan, and J. R. Anthis†
ACM SIGCHI Conference on Human Factors in Computing Systems

CHI 2025
Perceptions of Sentient AI and Other Digital Minds: Evidence from the AI, Morality, and Sentience (AIMS) Survey [DOI, arXiv] 🏅
J. R. Anthis, J. V. T. Pauketat, A. Ladak, and K. Manoli
ACM SIGCHI Conference on Human Factors in Computing Systems

CHI 2024
Which Artificial Intelligences do People Care Most About? A Conjoint Experiment on Moral Consideration [DOI, arXiv]
A. Ladak, J. Harris, and J. R. Anthis†
ACM SIGCHI Conference on Human Factors in Computing Systems

HRI 2024
A Taxonomy of Robot Autonomy for Human-Robot Interaction [DOI] 🏅
S. Kim, J. R. Anthis, and S. Sebo
ACM/IEEE International Conference on Human-Robot Interaction

NeurIPS 2023
Causal Context Connects Counterfactual Fairness to Robust Prediction and Group Fairness [DOI, arXiv]
J. R. Anthis and V. Veitch
Advances in Neural Information Processing Systems

JOURNAL PAPERS

2025 **Public Opinion and The Rise of Digital Minds: Perceived Risk, Trust, and Regulation Support** [DOI, arXiv]
J. Bullock, J. V. T. Pauketat, H. Huang, Y. Wang, and J. R. Anthis†
Public Policy and Management Review

2025 **World-Making for a Future with Sentient AI** [DOI (paywalled)], preprint]
J. V. T. Pauketat, A. Ladak, and J. R. Anthis†
British Journal of Social Psychology

2023 **Extending Perspective Taking to Nonhuman Animals and Artificial Entities** [DOI (paywalled), preprint]
A. Ladak, M. Wilks, and J. R. Anthis†
Social Cognition

2022 **Predicting the Moral Consideration of Artificial Intelligences** [DOI]
J. V. T. Pauketat and J. R. Anthis †
Computers in Human Behavior

2021 **The Moral Inclusion of Artificial Entities: A Literature Review** [DOI]
J. Harris and J. R. Anthis†
Science and Engineering Ethics

2021 **Moral Circle Expansion: A Promising Strategy to Impact the Far Future** [DOI]
J. R. Anthis and E. Paez
Futures

2020 **Institutional Change and the Limitations of Consumer Activism** [DOI]
J. R. Anthis
Humanities and Social Science Communications

2019 **Cell-Cultured Meat: Lessons from GMO Adoption and Resistance** [DOI]
J. Mohorčich and J. R. Anthis†
Appetite

WORKING PAPER

September 2025 **HumanAgencyBench: Scalable Evaluation of Human Agency Support in AI Assistants** [arXiv]
B. Sturgeon, D. Samuelson, J. Haimes, and J. R. Anthis

WORKSHOPS

ICLR 2025 **Human-AI Coevolution** [OpenReview]
J. R. Anthis, D. Asmar, K. R. Driggs-Campbell, A. Hardy, K. J. Meimandi, G. Keeling, M. Kochenderfer, H. Liu, R. Martín-Martín, A. A. Rushdi, M. R. Schlichting, P. Stone, H. Subramonyam, and D. Yang
International Conference on Learning Representations

CSCW 2024 **The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing** [DOI, arXiv]
A. Q. Zhang, R. Shaw, J. R. Anthis, A. Milton, E. Tseng, J. Suh, L. Ahmad, R. S. S. Kumar, J. Posada, B. Shestakofsky, S. T. Roberts, and M. Gray
ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing

EXTENDED ABSTRACTS

CSCW 2025 *(Forthcoming)* **"She's Like a Person but Better": Characterizing Emotional and Functional Relationships with Large Language Models**
K. Manoli, J. V. T. Pauketat and J. R. Anthis†
ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing

BOOK

2018 **The End of Animal Farming: How Scientists, Entrepreneurs, and Activists are Building an Animal-Free Food System**
J. R. Anthis
Beacon Press (Boston)

BOOK CHAPTERS

2023 **Animals**
J. R. Anthis and E. Soice
Modern Meat: The Next Generation of Meat From Cells, The Cellular Agriculture Society, Oxford University Research Archive (Oxford)

2023 **Consumers**
M. Wilks, J. R. Anthis, J. Y. Chung, and M. Vernon
Modern Meat: The Next Generation of Meat From Cells, The Cellular Agriculture Society, Oxford University Research Archive (Oxford)

2022 **Consciousness Semanticism: A Precise Eliminativist Theory of Consciousness**
J. R. Anthis
Biologically Inspired Cognitive Architectures 2021, eds. P. V. Klimov and D. J. Kelley, Springer (New York)

2017 **The Inner Lives of Farmed Animals**
J. R. Anthis and P. Singer
The Reducetarian Solution, ed. B. Kateman, TarcherPerigree (New York)

GRANTS AND AWARDS

2025 **Rising Stars in Data Science**
Stanford Data Science in collaboration with UC San Diego and University of Chicago

2025 **University of Chicago Sociology Dissertation Fellowship**
Award amount of $12,000

2016–Current **Sentience Institute**
Funding of over $1,500,000 as principal investigator of 501(c)(3) nonprofit research lab

CHI 2025 **Best Paper Honorable Mention** 🏅
Awarded to J. R. Anthis et al. for "Perceptions of Sentient AI and Other Digital Minds: Evidence From the AI, Morality, and Sentience (AIMS) Survey"

HRI 2024 **Best Paper Honorable Mention** 🏅
Awarded to S. Kim et al. for "Taxonomy of Robot Autonomy for Human-Robot Interaction"

2020 **University of Chicago PhD Fellowship**
Full funding and living stipend for PhD studies, equivalent to approximately $600,000, including tuition. Numerous competitive travel grants from various sources.

PRESENTATIONS

2025 **HumanAgencyBench: Measuring AI Threats to Human Agency**
Center for Human-Compatible AI Annual Workshop (June)
Google DeepMind (May)

2025 **Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation**
Human-Centered Evaluation and Auditing of Language Models (HEAL), workshop at CHI 2025

2024 **What Do Red Teamers Do?**
The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing, workshop at CSCW 2024

2024 **Framing Contests: How Human-Computer Interaction Shapes the Future of AI**
From Stem to Stern: Contestability Along AI Value Chains, workshop at CSCW 2024

2024 **The AI Double Standard: Humans Judge AIs More Harshly for One Bad AI Than They Judge Humans for One Bad Human**
Trust and Reliance in Human-AI Workflows (TREW), workshop at CHI 2024

2024 **The Impossibility of Fair LLMs**
Human-Centered Evaluation and Auditing of Language Models (HEAL), workshop at CHI 2025

2023 **Morality is a Two-Way Street: The Role of Mind Perception and Moral Attribution in AI Safety**
AI meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics (MP2), workshop at NeurIPS 2023

2023 **Testing the Limits of Chatbots** [video]
Cypher (prompt injection workshop with attack/defense at India's largest AI conference)

2023 **The Rise of Digital Minds** [video]
Cypher

2023 **Historicism in Human-AI Interaction: AI as a General Purpose Technology**
Historicism in/as CSCW Method: Research, Sensibilities, and Design, workshop at CSCW 2023

2022 **Framing a General Purpose Technology: The Case of AI Multiobjectivity**
Strategic Management Society Annual Conference

2022 **AI Safety Needs Data Scientists**
Imperial College London

2022 **Field-Level Framing Contests in Artificial Intelligence**
AI and Strategy Consortium

2021 **The Individual-Institutional Gap: Evidence from the Animals, Food, and Technology (AFT) Survey**
Annual Meeting of the American Sociological Association

2021 **The Many Frames of Artificial Intelligence: Preliminary Grounded Theory**
European Group for Organizational Studies Colloquium

2021 **Emergent Trends in Public Opinion on Animal Farming and Animal Product Alternatives**
American Association for the Advancement of Science Annual Meeting

2020 **Institutional Change and the Limitations of Consumer Activism**
Annual Meeting of the American Sociological Association

2020 **Moral Circle Expansion**
Annual Meeting of the American Sociological Association

2018 **Animal Ethics and Moral Circle Expansion** [video]
Effective Altruism Global

2017 **When Is Confrontation Effective?**
Animal Rights National Conference

2016 **Wild Animal Suffering**
International Animal Rights Conference

2016 **Is Individual Consumption an Effective Tool for Social Change?**
Oxford University

2015–2023 **Effective Altruism**
Presented more than 30 times as a general introduction to ideas such as charity program evaluation, moral circle expansion, and impact-focused career choice.

2016–2019 **The End of Animal Farming**
Book presentation delivered more than 40 times at U.S. institutions including Harvard, Penn, Northwestern, and Yale as well as institutions in 15 other countries. Short version delivered in February 2018 at TEDxUniversityOfMississippi with over 200,000 views and selected as a small number of TEDx talks featured on TED.com

DATASETS AND RESOURCES

2021–Current **Artificial Intelligence, Morality, and Sentience (AIMS) Survey** [link]
J. V. T. Pauketat, A. Ladak, and J. R. Anthis
Sentience Institute

2017–Current **Animals, Food, and Technology (AFT) Survey** [link]
J. R. Anthis and A. Ladak
Sentience Institute

2017–Current **Global Farmed Animal Estimates** [link]
K. Anthis and J. R. Anthis
Sentience Institute

2017–Current **US Factory Farming Estimates** [link]
J. R. Anthis
Sentience Institute

TEACHING

Spring 2023 **Hierarchical Linear Models**
Teaching Assistant
Instructor: Stephen Raudenbush

Fall 2022 **Business Statistics**
Teaching Assistant
Instructor: Bryon Aragam

Winter 2022 **Computational Content Analysis**
Teaching Assistant
Instructor: James Evans

SELECTED PUBLIC ESSAYS

2025 **It's Time to Prepare for AI Personhood** [link]
The Guardian

2023 **Why We Need a "Manhattan Project" for A.I. Safety** [link]
Salon

2023 **We Need an AI Rights Movement** [link]
The Hill

2020 **Utilitarianism in an Age of Moral Turbulence** [link]
Arc Digital

2019 **A Foie Gras Ban is Overdue** [link]
The Guardian

2019 **China and India Could Lead the Way in Adopting 'Clean Meat'** [link]
The Guardian

2018 **'Clean Meat,' the Future of Vegetarianism** [link]
National Review

2018 **Is "Clean Meat" the Solution to Industrial Animal Farming?** [link]
Georgetown Journal of International Affairs

2018 **Our Treatment of Animals Is Stalling Human Progress** [link]
Quartz

2016 **The Moral Demand for Cultured Meat** [link]
Salon

2015 **Wild Animals Endure Illness, Injury, and Starvation. We Should Help.** [link]
Vox

SERVICE

Area chair / associate chair / metareviewer:

CHI (2026)

CHI Late Breaking Work track (2025)

CSCW (2025 July cycle, October cycle)

Secondary area chair:

CSCW (2026)

Peer reviewer (conferences):

AAAI (2025, 2026): Association for the Advancement of Artificial Intelligence

ACL (2025): Association for Computational Linguistics Rolling Review

CHI (2024, 2025): ACM Conference on Human Factors in Computing Systems

CSCW (2024): ACM Conference on Computer-Supported Cooperative Work

DIS (2025): ACM Conference on Designing Interactive Systems

ICLR (2025): International Conference on Learning Representations

NeurIPS (2024, 2025): Advances in Neural Information Processing Systems

Peer reviewer (journals): *AI and Ethics*, *American Journal of Sociology*, *Artificial Intelligence Review*, *Digital Society*, *Frontiers in Artificial Intelligence*, *Frontiers in Sustainable Food Systems*, *New Techno-Humanities*, *Societal Impacts*

Peer reviewer (workshops): AI Meets Moral Philosophy and Moral Psychology (MP2, NeurIPS 2023), Human-Centered Evaluation and Auditing of Language Models (HEAL, CHI 2025, 2026), Trust and Reliance in Evolving Human-AI Workflows (TREW, CHI 2025)

MISCELLANEOUS

Red team. OpenAI. 2023 to present.

Affiliate. Center for Human-Compatible Artificial Intelligence (CHAI), University of California, Berkeley. 2023 to present.

Research covered by global media, e.g., *The Atlantic*, *New Scientist*, *Forbes*, and *Vox*.

SIGCHI Emerging Scholars. 2023. Minneapolis, USA.

Lead organizer of Summit on AI in Society (meeting of top AI scholars). 2022. Chicago, USA.

Diverse Intelligences Summer Institute. 2022. St. Andrews, UK.

Summer Institute in Computational Social Science. 2021. Princeton, USA (virtual).

App Academy (web development bootcamp). 2015. San Francisco, USA.