KATHERINE L. ELKINS, Ph.D.

Professor of Humanities & Comparative Literature · Kenyon College
Director, Integrated Program in Humane Studies · Founding Co-Director, AI CoLab / Human-Centered AI Lab
elkinsk@kenyon.edu · katherineelkins.com · ORCID 0000-0001-9887-4854

Katherine Elkins is a researcher in AI safety, governance, and the public interest, working where the humanities meet computing and public policy. She leads the Modern Language Association's team at the U.S. AI Safety Institute / NIST AI Consortium (now CAISI), bringing humanities expertise to national AI safety standards, and serves as Principal Investigator of Schmidt Sciences' Archival Intelligence project, building open-source, community-governed AI for endangered cultural archives. Her parallel work spans AI as public infrastructure and international cultural-heritage policy at UNESCO. In 2016 she co-created the first human-centered AI curriculum, and she went on to develop the first ensemble-based method for diachronic sentiment analysis of narrative, to fine-tune and evaluate large language models early (GPT-2, 2019), and to publish the first study testing whether an AI system could pass as a human author (2020). That early empirical work now grounds a sustained, critical inquiry into whether and how AI can advance the public good.

AI GOVERNANCE, PUBLIC INFRASTRUCTURE & CULTURAL HERITAGE

Lead, Modern Language Association team, U.S. AI Safety Institute / NIST AI Consortium (now CAISI) — humanities expertise for national AI safety standards; team includes Jon Chun, Rita Raley, Seth Perlow, and Anna Mills (2024–present).

Principal Investigator, Schmidt Sciences HAVI — "Archival Intelligence: Rescuing New Orleans' Endangered Cultural Legacy," open-source, community-governed AI for endangered cultural archives (2025–present).

UNESCO AI & Cultural Heritage Initiative (2024–present); UNESCO CSO Network on the Ethics of AI, Subgroup on Intellectual Property and Culture (2026).

Public AI — international network for AI as public infrastructure; co-author of open-source / Public AI position papers presented at ICML venues (2023–present).

Meta Open Innovation AI Research Community (London) — Open Source Working Group, later Transparency Working Group (2023–2026).

Co-Founder and Co-Director, Human-Centered AI Lab, Inc. (Ohio nonprofit), with Jon Chun.

OpenAI Higher Education Forum — Education Guild selected speaker, 1 of 6, invitation-only in-person session on leveraging AI for humanities and social science research (San Francisco, 2025).

Steering Committee, ADHO Computational Literary Studies (2026); MLA AI & Research Working Group; MLA MAPS Leadership Institute (2025); Advisory Board, The Helix Center (2022–present).

SELECTED GRANTS & HONORS

Schmidt Sciences HAVI Award — one of 23 worldwide; up to $330,000 over 18 months. Principal Investigator, "Archival Intelligence: Rescuing New Orleans' Endangered Cultural Legacy" (2025).

Notre Dame–IBM Tech Ethics Lab Grant — competitive international award. Principal Investigator, ethics-based auditing of leading LLMs for fairness, accuracy, transparency, and explainability (2024).

NEH Distinguished Teaching Professorship, Kenyon College (2020–2023).

Whiting Fellowship (Mrs. Giles Whiting Foundation).

A. Owen Aldridge Prize in Comparative Literature.

Kenyon Senior Faculty Trustee Teaching Excellence Award.

BOOKS & EDITED VOLUMES

The Shapes of Stories: Sentiment Analysis for Narrative. Cambridge Elements in Digital Literary Studies. Cambridge University Press, 2022.

Proust's In Search of Lost Time: Philosophical Perspectives. Oxford Studies in Philosophy and Literature. Oxford University Press, 2022. Editor and contributor.

Guest Editor, "Depiction of Good and Evil in Fairytales," *Humanities* (MDPI), special issue, 2024–25.

SELECTED PUBLICATIONS & POLICY RESEARCH

AI, Cultural Heritage & the Public Good

"Provenance Infrastructure as a Safeguard for Cultural Commons in the Age of Generative AI" (with J. Chun). Selected contribution, UNESCO Subgroup on Intellectual Property and Culture, 2026.

"The Dual Edge: AI's Impact on Cultural Heritage and Preservation." UNESCO. Forthcoming.

"Position: If open source is to win, it must go public" (with J. Tan, N. Vincent, M. Sahlgren, J. Low, D. Pham, S. Pyysalo, and J. Jitsev). Accepted, *Proceedings of ICML* 2026 (Position Paper track). Earlier version presented as a spotlight talk at CODEML@ICML 2025.

"Position: Near to Mid-term Risks and Opportunities of Open-Source Generative AI." *Proceedings of ICML* 2024. Oral presentation, top ~2%; international collaboration across Oxford, UC Berkeley, ITS Rio, and Bocconi.

AI Governance & Regulation

"The AI Governance Trilemma" (with C. Schroeder de Witt and J. Chun). Under review, 2026.

"Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US" (with C. Schroeder de Witt and J. Chun). 2024.

"Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit" (with J. Chun). arXiv, 2024.

AI, Culture & the Humanities

"Can We Fall in Love with AI Fiction? The AI Fiction Paradox." *Modern Fiction Studies*. Forthcoming 2027.

"A(I) University in Ruins: What Remains in a World with Large Language Models?" *PMLA*, 2024.

"AI Comes for the Author." *Poetics Today* 45, no. 2, 2024.

"Can GPT-3 Pass a Writer's Turing Test?" (with J. Chun). *Journal of Cultural Analytics* 5, no. 2, 2020.

"In Search of a Translator: Using AI to Evaluate What's Lost in Translation." *Frontiers in Computer Science* 6, 2024.

"Generative AI and the Stakes of Literary Form." In *The New Literature and AI Studies*, ed. Robert Marzec. Cambridge University Press. Forthcoming.

Computational Methods & AI Safety

"Syntactic Framing Fragility: An Audit of Robustness in LLM Ethical Decisions" (with J. Chun). arXiv:2601.09724; submitted, FAccT 2026.

"When Prohibitions Become Permissions: Auditing Negation Sensitivity in Language Models" (with J. Chun). arXiv:2601.21433, 2026.

"AgenticSimLaw: A Juvenile Courtroom Multi-Agent Debate Simulation" (with J. Chun and Y. S. Lee). AAAI-26 LaMAS.

Earlier scholarship on Plato, Baudelaire, Proust, Kafka, Wordsworth, and Maryse Condé appears in *Philosophy and Literature, Comparative Literature Studies, MLN, Discourse, Modern Language Quarterly*, and volumes from Springer and De Gruyter.

STANDARDS, POLICY CONTRIBUTIONS & RESEARCH IMPACT

As lead of the Modern Language Association team in the NIST AI Consortium (CAISI) since its 2024 founding, authored three public comments to NIST in 2026: the Agent Security RFI (NIST-2025-0035), introducing interpretive tractability; the AI 800-2 benchmarking framework; and the NCCoE AI agent identity and authorization concept paper. Findings from the first comparative ethics-based audit of deployed LLMs were presented by invitation at Meta's London headquarters and to U.S. AI Safety Institute Working Group 5 / Safety; the work was later highlighted at the consortium's inaugural open plenary.

Field-defining contributions include: first human-centered AI curriculum (2016); first ensemble-based method for diachronic sentiment analysis of narrative; first study testing whether an AI system could write convincingly as a human author (GPT-3, 2020); first comparative ethics-based audit of deployed LLMs; first operational Syntactic Framing Fragility / Syntactic Variation Index method.

A cross-disciplinary method: the diachronic sentiment-analysis method (SentimentArcs is one implementation) is now infrastructure across cognitive science, NLP, AI visualization, and digital humanities, with adoption in NarraBench (EACL 2026), Story Ribbons (IEEE TVCG 2026), a 25,728-retelling study in Scientific Reports (2023), Litteraturmaskinen (ACL 2026), Digital Scholarship in the Humanities (2025), political theory, and sustainability research.

AI evaluation & governance reception: the 2020 GPT-3 study is a canonical case in machine-authorship debates (Floridi and Chiriatti, Minds and Machines, 2020); the ethics-based LLM audit was applied at scale in COLING 2025 and named among foundational works in an AAAI 2026 alignment survey; the ICML 2024 open-source position paper is cited among canonical openness frameworks at FAccT 2025.

Recognition: the human-centered AI curriculum is cited by Fields Medalist Terence Tao and Tanya Klowden in their essay on mathematical methods and human thought in the age of AI.

Global reach: 400+ mentored student research projects; more than 107,000 downloads reaching 198 countries.

PUBLIC ENGAGEMENT & PUBLIC-FACING SCHOLARSHIP

Press and public analysis

Christian Science Monitor: "As AI Leaps Forward, Concern Rises That Innovation Is Leaving Safety Behind" — quoted as a NIST AI Safety Institute Consortium contributor (2026).

NPR / WOSU: "Could Artificial Intelligence Save Endangered Archives?" — lead subject, Schmidt Sciences HAVI / Archival Intelligence (2026).

Forbes: "Where AI Meets the Humanities: Inside Kenyon College's Bold Experiment" (2025).

Engineering (Chinese Academy of Engineering / Elsevier): "AI's Talent for Translation Lowers Language Barriers" (2025).

Al Jazeera The Stream: "Is AI Better at Making Art Than Humans?" debate (2023).

OpenAI, Bloomberg, and public research translation

OpenAI Higher Education Forum, Education Guild selected speaker, invitation-only in-person event, San Francisco (2025).

OpenAI Forum virtual event, "Discoveries Across Disciplines" / "AI and Academic Research," with Kevin Weil (VP, OpenAI for Science) and Leonardo Impett (2025).

Bloomberg / Emeritus: AI Strategy Course, on-screen subject-matter expert (2024). MLA Member Spotlight (2025).

Audio, podcasts, and long-form public humanities

Humanity at Scale: "From Homer to GPT: The Collision of Human Imagination and AI" (2025); *Creative Velocity*: "AI Deep Dive" (2026).

Merging Minds (BureauWorks): "Translating Worlds" (2024); *RadioAI*: "ChatGPT and Large Language Models" (2023).

Helix Center roundtables, New York: "Are Natural Language Generators for Real?" (2022) and "Emotion" (2023).

Audible / The Modern Scholar lecture series: *The Modern Novel; The Giants of French Literature; Odyssey of the West VI*.

SELECTED KEYNOTES & DISTINGUISHED LECTURES

Invited talk on the dual edge of AI's impact on culture, UNESCO Cairo office (2025).

Keynote, "Governing AI in Partial View: Why Responsible AI Needs the Humanities," University of Michigan MIDAS Annual Data Science and AI Summit (accepted, December 2026).

METC Conference keynote, Weill Cornell Medicine–Qatar, Doha (CME credits, 2025).

Opening Plenary, Faith, Reason and World Affairs Symposium, Concordia College (2025).

Keynote, WPI Global Lab / Global School, "AI at a Crossroads: Who Shapes the Future?" (2025).

Kahn Liberal Arts Institute, Smith College (2025).

"Beyond AI Literacy," Dartmouth (2024).

Northrup Distinguished Lecture, Yale University (2024).

Day of Digital Humanities keynote, Carleton College (2024).

AI Literacy Across the Curriculum, Lafayette College (2024).

A. J. Carlson Lecture, Austin College (2024).

Meredith Donovan Lecture, Mount St. Mary's University (2023).

AI Working Group featured lecture, Wofford College (2023).

"Why Can't We Fall in Love with AI Fiction?" International Society for the Study of Narrative, Aarhus (June 2026).

"Programming Humanity," Ohio State University (2019).

"Stories that Win" Symposium, Washington University in St. Louis (2024).

SELECTED INVITED TALKS & PANELS

"Civic Education After AI" / AI and Democracy, Ohio State University, Chase Center (April 2026).

Invited participant, Schmidt Sciences HAVI 2026 Convening, London (October 2026).

"AI and Fiction" panel (Prose Fiction forum), MLA Convention, Los Angeles (January 2027).

Chronicle of Higher Education Virtual Forum, invited speaker (2025).

RALLY Innovation Conference, "Human Algorithms," Indianapolis (2025).

Yale Alumni in AI (2025).

Helix Center Roundtables, New York — with Ned Block, Francesca Rossi, Joseph LeDoux, Rosalind Picard, and Kyunghyun Cho (2022–23).

"AI Needs the Humanities (and the Humanities Need AI)," University of Tennessee, Knoxville (2023).

"Human-Machine Decision Making and the Convergence of Intelligence," McGill / Pittsburgh Supercomputing Center / Carnegie Mellon (2023).

First transdisciplinary AI research presentation, Modernist Studies Association (2019).

CREATIVE PRACTICE & RESIDENCIES

Artist-in-Residence, WPI Global Lab, Worcester Polytechnic Institute (2025).

DivaBot (2021) — transformer-based AI performing live improvisation, with Jon Chun and Jim Dennen, for the centennial of Čapek's *R.U.R.*, the play that introduced the word "robot."

TEACHING & TECHNICAL BACKGROUND

Curriculum leadership: co-created the first human-centered AI curriculum (2016) and the AI CoLab with Jon Chun; 400+ mentored student research projects, with students moving into computer science, law, data science, AI research, product, and governance roles, and public-interest technology work. Selected by graduating seniors to deliver the 2024 Baccalaureate Address.

Selected courses: Programming Humanity; AI for Humanity; Cultural Analytics; Senior Research Seminar; courses combine Python, data visualization, probabilistic methods, Bayesian statistics, deep neural networks, transformers, LLMs, ethics, surveillance, privacy, narrative, and cultural analysis.

Technical background: hands-on work since 2019 spans fine-tuning and evaluation across GPT-2 to current models (Python, PyTorch, HuggingFace), red-teaming, explainable AI, multi-agent simulation, and the Syntactic Framing Fragility / Syntactic Variation Index method.

SELECT REVIEWING

ICML, NeurIPS, AAAI/AIES, PNAS, Science Advances, Journal of Cultural Analytics, ACM Journal on Computing and Cultural Heritage, and Oxford University Press; grant and program review for the National Endowment for the Humanities, the Swiss National Science Foundation, and Schmidt Sciences (Trustworthy AI).

EDUCATION

Ph.D., Comparative Literature. University of California, Berkeley.

B.A., Literature (Distinction in Major, cum laude). Yale University.

LANGUAGES

French, German, Spanish (professional proficiency); **Latin, Ancient Greek** (reading proficiency).