Rebecca L. Johnson

PhD in AI Ethics
MA (Research), B.Sc., B.A.

Ethics, evaluation, and human-centred governance of generative AI

CONTACT

rebecca.johnson@sydney.edu.au

The University of Sydney,
Australia (AEST)

www.EthicsGenAI.com

www.EthicsGenAI.com/media

linkedin.com/in/becjohnson

CORE STRENGTHS

- Ethics of AI models
- Responsible Evaluation of AI
- Sociotechnical Risk Mapping for transparency and accountability
- Interdisciplinary collaboration
- AI Evaluation frameworks
- Academic & Professional writing
- Teaching and curriculum design
- Public communication of AI ethics
- Teamwork & interpersonal skills
- Collaborative and empathic leadership
- Research and project design
- Workshop facilitation
- Conference design and hosting
- Public speaking

AWARDS

- Stanford HAI Scholarship
- Postgrad Research Prize for Leadership - Uni. Sydney
- Australian Government Research Training Stipend
- MIT EmTech Scholarships (2019, 2020)
- Isabel Paulette Postgraduate Careers Scholarship

PROFILE

Researcher and educator in the ethics and governance of generative AI. I develop evaluation methods and sociotechnical approaches that keep human values at the centre of AI design and use. Ethics, to me, is not a checklist but an ethos embedded in how technologies are imagined, built, and applied. With foundations in both Science and the Arts, I work across disciplines to help people understand these technologies in a human context. Skilled at clear, engaging teaching that makes complex ideas accessible and connects ethical principles to real-world AI practice.

SELECTED EXPERIENCE

The University of Sydney 2016 - PRESENT

Educator - Faculties of Science, Arts, and Business

- Designed and delivered units on Ethics in Science and Technology, Leadership in Business, Organisational Communications, Media & Communications, and Philosophy of Science;
- Coached new tutors;
- Student feedback typically 4.6–4.9 out of 5 across large cohorts.

Google Research 2021 - 2022

Researcher - The Ethical AI Team

- Ran applied ethics experiments with LLMs (LaMDA, PaLM); contributed to responsible prompt design and evaluation design.
- Developed *The World Values Benchmark* methodology to describe model behaviour relative to human population value profiles

MOST RECENT EDUCATION

PhD - AI Ethics 2019 - 2025
Philosophy of Science | University of Sydney
Under examination

Master's by Research - Communications 2016 - 2018
Media & Communications| University of Sydney
Awarded with Distinction

SELECTED PUBLICATIONS

The Ghost in the Machine Has an American Accent: Value Conflict in GPT-3 - Preprint 2022 arXiv:2203.07785, updated version currently under journal review.

What are AI Researchers Really Arguing About- in "Handbook on the Ethics of Artificial Intelligence", Edward Elgar Publishing, 2024