Hannah Rose Kirk
hannah.kirk@oii.ox.ac.uk
hannahrosekirk.com

Education

2020–2026 **DPhil**, Social Data Science, Oxford Internet Institute, University of Oxford
- Full PhD Scholarship from ESRC Grand Union DTP
- **DPhil Thesis:** *The Empirical Dynamics of Human-AI Alignment*

2020–2021 **MSc**, Social Data Science, Oxford Internet Institute, University of Oxford
Grade: Distinction (77%)
- Awarded the Oxford Internet Institute Thesis Prize for the best MSc thesis (2021)
- **MSc Thesis:** *Hatemoji: The Construction and Classification of Emoji-based Hate Speech*

2018–2020 **MA**, Economics and China Studies, Yenching Academy, Peking University
GPA: 3.99, Rank: 2/99
- Awarded Robin Li Scholarship for exceptional students in technical fields
- Outstanding Academic Prize and Outstanding Research Prize
- **Thesis:** *Context and Culture Aware Recommender Systems in Chinese Society*

2015–2018 **BA**, Economics, Trinity College, University of Cambridge
Double First Class Honours
- Junior Scholar (2016-17), Senior Scholar (2017-18)
- Trinity Dennis Robertson Prize for best undergraduate dissertation (2018)
- Trinity Roger Moore Prize in Part I of the Economics Tripos (2016)
- **Thesis:** *Cooperation and Creed: An Experimental Study of Religiosity*

Professional Experience

2025–now Research Workstream Lead, UK AI Security Institute, London, UK
Lead a team of 10+ research scientists to deliver high quality evidence on human influence risks of frontier AI for policymakers, academia, and developer labs

2025–now Postdoctoral Research, Department of Experimental Psychology, Oxford University
Research on human-AI value alignment, and over-reliance of AI in learning

2024–2025 Research Scientist, UK AI Security Institute, London, UK
Conducted research (RCTs) on social and psychological capabilities of frontier AI to build world-leading evaluations within the UK government

2023 Visiting Academic, New York University, New York City, USA
Researched LLM alignment and human-AI collaboration, hosted by Prof. He He's & Prof. Sam Bowman's labs

2023 Consultant, OpenAI, Remote
Consulted on strategy for responsible AI pathways and assessed model safety via pre-launch adversarial testing

2022–2023 External Student Research Collaborator, Google Research, Remote

Designed adversarial data-centric challenge to identify unsafe failure modes in text2image models (FAccT-2024)

2021–2023 Data Scientist in Online Safety, The Alan Turing Institute, London, UK
Led development of data-centric AI for abuse detection; delivered report on footballer abuse featured in 30+ countries

2021–2023 Research Scientist, Rewire, London, UK
Co-developed product and engineering solutions to build socially responsible AI for online safety

2020–2023 Research Labs Principal Investigator, Oxford AI Society, Oxford, UK
Led graduate student teams to publish six papers in leading conferences (NeurIPs, ACL)

2019–2020 Research Intern, The Berggruen Institute, China Center, Beijing, China
Led and published research on cultural philosophies in artificial intelligence & digital ethics

Awards & Honors

2024 **Best Paper Award**, NeurIPS 2024 [C2]
2024 **Two Oral Presentations, NeurIPS 2024, Top 0.5% of papers** [C2, C3]
2024 **Outstanding Paper Award**, ACL 2024 [C6]
2023 **Best Paper Award**, SemEval @ ACL 2023 [W5]
2021 **Oxford Internet Institute Thesis Prize** for the best MSc thesis [C14], University of Oxford
2019 **Robin Li Scholarship** for exceptional students in technical fields
2019 **Yenching Academy Prize for Outsanding Academic Performance**
2019 **Yenching Academy Prize for Outstanding Field Research**
2018 **Trinity Dennis Robertson Prize** for best undergraduate dissertation [P14], Trinity College, University of Cambridge
2016 **Trinity Roger Moore Prize** in Part I of the Economics Tripos, Trinity College, University of Cambridge

Grants & Funding

2023–2024 Microsoft Research Accelerating Foundation Models Grant ($40,000)
Personalised and diverse feedback for humans-and-models-in-the-loop

2022–2024 Meta AI Dynabench Data Collection and Benchmarking Platform Grant ($47,000)
Optimising feedback between humans-and-model-in-the-loop

2020–2024 **Fully-Funded PhD Scholarship** from the Economic and Social Science Research Council
Digital Social Science pathway, ES/P000649/1

Publications

Google Scholar, citations = 4255, h-index = 28, i10-index = 36
♥ First authorship ♡ Joint first authorship ♣ Senior authorship ♧ Joint senior authorship

Journal Articles

J1. Bean, A. M., Payne, R. E., Parsons, G., **Kirk**, H. R., Ciro, J., Mosquera, R., Hincapié M, S., Ekanayaka, A. S., Tarassenko, L., Rocher, L., Mahdi, A., *et al.* Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine* (2026).

J2. Castro-Gonzales, L., Chung, Y.-L., **Kirk**, H. R., Francis, J., Williams, A., Johansson, P. & Bright, J. Cheap Learning: Maximising Performance of Language Models for Social Data Science Using Minimal Data. *Sociological Methods and Research* (2025).

J3. Don-Yehiya, S., Burtenshaw, B., Astudillo, R. F., Osborne, C., Jaiswal, M., Kuo, T.-S., Zhao, W., Shenfeld, I., Peng, A., Yurochkin, M., **Kirk**♔, H. R., *et al.* The Future of Open Human Feedback. *Nature Machine Intelligence* (2025).

J4. **Kirk**♥, H. R., Gabriel, I., Summerfield, C., Vidgen, B. & Hale, S. A. Why Human–AI Relationships Need Socioaffective Alignment. *Nature Humanities and Social Sciences Communications* **12,** 1–9. ISSN: 2662-9992 (May 2025).

J5. **Kirk**♥, H. R., Vidgen, B., Röttger, P. & Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* **6,** 383–392 (2024).

J6. Mökander, J., Schuett, J., **Kirk**, H. R. & Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics* **4,** 1085–1115 (2024).

J7. Osborne, C., Ding, J. & **Kirk**, H. R. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. *Journal of Computational Social Science* **7,** 2067–2105 (2024).

J8. **Kirk**♡, H. R. & Gupta, S. The mediation of matchmaking: a comparative study of gender and generational preference in online dating websites and offline blind date markets in Chengdu. *The Journal of Chinese Sociology* **9,** 2 (2022).

J9. **Kirk**♥, H. R., Lee, K. & Micallef, C. The nuances of Confucianism in technology policy: An inquiry into the interaction between cultural and political systems in Chinese digital ethics. *International Journal of Politics, Culture, and Society*, 1–24 (2020).

Peer-reviewed Conference Proceedings

C1. Christian, B., **Kirk**, H. R., Thompson, J., Summmerfield, C. & Dumbalska, T. *Reward Model Interpretability via Optimal and Pessimal Tokens* in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025).

C2. **Kirk**♥, H. R., Whitefield, A., Rottger, P., Bean, A. M., Margatina, K., Mosquera-Gomez, R., Ciro, J., Bartolo, M., Williams, A., He, H., *et al.* *The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models* in *Advances in Neural Information Processing Systems* (2025).
– Best Paper Award
– Oral Presentation, Top 0.5% of papers.

C3. Bean, A. M., Hellsten, S., Mayne, H., Magomere, J., Chi, E. A., Chi, R., Hale, S. A. & **Kirk**♣, H. R. *Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages* in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2024).

– **Oral Presentation, Top 0.5% of papers**.

C4. Khandelwal, K., Tonneau, M., Bean, A. M., **Kirk**, H. R. & Hale, S. A. *Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models* in *Proceedings of the 2024 International Conference on Information Technology for Social Good (GoodIT '24)* (Association for Computing Machinery, 2024), 231–239.

C5. Quaye, J., Parrish, A., Inel, O., Rastogi, C., **Kirk**, H. R., Kahng, M., Van Liemt, E., Bartolo, M., Tsang, J., White, J., *et al. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation* in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024), 388–406.

C6. Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., **Kirk**, H. R., Schütze, H. & Hovy, D. *Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models* in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024).
– **Outstanding Paper Award**.

C7. Hall, S. M., Gonçalves Abrantes, F., Zhu, H., Sodunke, G., Shtedritski, A. & **Kirk**♣, H. R. *VISOGENDER: A dataset for benchmarking gender bias in image-text pronoun resolution* in *Advances in Neural Information Processing Systems* (2023).

C8. **Kirk**♥, H. R., Bean, A. M., Vidgen, B., Röttger, P. & Hale, S. A. *The past, present and better future of feedback learning in large language models for subjective human preferences and values* in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023).

C9. Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., Diamos, G., He, L., Parrish, A., **Kirk**, H. R., *et al. Dataperf: Benchmarks for data-centric ai development* in *Advances in Neural Information Processing Systems* (2023).

C10. Röttger, P., **Kirk**, H. R., Vidgen, B., Attanasio, G., Bianchi, F. & Hovy, D. *XSTEST: A test suite for identifying exaggerated safety behaviours in large language models* in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (2023).

C11. Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., **Kirk**♧, H. R., Shtedritski, A. & Bain, M. *A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning* in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (2022).

C12. **Kirk**♥, H. R., Birhane, A., Vidgen, B. & Derczynski, L. *Handling and Presenting Harmful Text in NLP* in *Findings of the Association for Computational Linguistics: EMNLP 2022* (2022).

C13. **Kirk**♥, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A. & Asano, Y. *Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models* in *Advances in neural information processing systems* (2021).

C14. **Kirk**♥, H. R., Vidgen, B., Röttger, P., Thrush, T. & Hale, S. A. *Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate* in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics* (2021).

Peer-reviewed Workshop Proceedings

W1. Padmakumar, V., Jin, C., **Kirk**♡, H. R. & He, H. *Beyond the Binary: Capturing Diverse Preferences With Reward Regularization* in *Socially Responsible Language Modelling Research Workshop, NeurIPs 2024* (2024).

W2. Quaye, J., Parrish, A., Inel, O., Kahng, M., Rastogi, C., **Kirk**, H. R., Tsang, J., Clement, N. L., Mosquera, R., Ciro, J. M., *et al. Lexically-constrained automated prompt augmentation: A case study using adversarial T2I data* in *NeurIPs Safe Generative AI Workshop 2024* (2024).

W3. Williams, A. R., **Kirk**, H. R., Burke-Moore, L., Chung, Y.-L., Debono, I., Johansson, P., Stevens, F., Bright, J. & Hale, S. A. *DoDo Learning: Domain-Demographic Transfer in Language Models for Detecting Abuse Targeted at Public Figures* in *Proceedings of the Fourth Workshop on Threat, Aggression and Cyberbullying (LREC-COLING-2024)* (2024), 134–154.

W4. **Kirk**♥, H. R., Vidgen, B., Röttger, P. & Hale, S. A. *The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models* in *Socially Responsible Language Modelling Research Workshop, NeurIPs 2023* (2023).

W5. **Kirk**♥, H. R., Yin, W., Vidgen, B. & Röttger, P. *Semeval-2023 task 10: Explainable detection of online sexism* in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), ACL* (2023).
– **Best Paper Award**.

W6. Smith, B., Farinha, M., Hall, S. M., **Kirk**♲, H. R., Shtedritski, A. & Bain, M. *Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets* in *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI* (2023).

W7. Borchers, C., Gala, D. S., Gilburt, B., Oravkin, E., Bounsi, W., Asano, Y. M. & **Kirk**♣, H. R. *Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements* in *Proceedings of the 4th workshop on gender bias in natural language processing (NAACL)* (2022).

W8. **Kirk**♥, H., Vidgen, B. & Hale, S. A. *Is More Data Better? Re-thinking the Importance of Efficiency in Abusive Language Detection with Transformers-Based Active Learning* in *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)* (2022).

W9. **Kirk**♥, H. R., Jun, Y., Rauba, P., Wachtel, G., Li, R., Bai, X., Broestl, N., Doff-Sotta, M., Shtedritski, A. & Asano, Y. M. *Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset* in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (2021).

Working papers and pre-prints

P1. Gausen, A., Wallbridge, S., **Kirk**, H. R., Williams, J. & Summerfield, C. Disclosure By Design: Identity Transparency as a Behavioural Property of Conversational AI Models. *arXiv preprint arXiv:2603.16874* (2026).

P2. **Kirk**♥, H. R., Davidson, H., Saunders, E., Luettgau, L., Vidgen, B., Hale, S. A. & Summerfield, C. Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv preprint arXiv:2512.01991* (2026).

P3. Bean, A. M., Kearns, R., Romanou, A., Hafner, F., Mayne, H., **Kirk**, H. R., *et al.* Measuring what Matters: Construct Validity in Large Language Model Benchmarks (2025).

P4. Luettgau, L., Cheung, V., Dubois, M., Juechems, K., Bergs, J., Davidson, H., O'Dell, B., **Kirk**, H. R., Rollwage, M. & Summerfield, C. People readily follow personal advice from AI but it does not improve their well-being. *arXiv preprint arXiv:2511.15352* (2025).

P5. Luettgau, L., **Kirk**, H. R., Hackenburg, K., Bergs, J., Davidson, H., Ogden, H., Siddarth, D., Huang, S. & Summerfield, C. Conversational AI increases political knowledge as effectively as self-directed internet search. *arXiv preprint arXiv:2509.05219* (2025).

P6. Ryström, J., **Kirk**, H. R. & Hale, S. Multilingual!= Multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. *arXiv preprint arXiv:2502.16534* (2025).

P7. Shah, A., South, T., Evans, T., **Kirk**, H. R., Trask, A., Weyl, E. G. & Bakker, M. A. Robust AI Personalization Will Require a Human Context Protocol (2025).

P8. Summerfield, C., Luettgau, L., Dubois, M., **Kirk**, H. R., Hackenburg, K., Fist, C., Slama, K., Ding, N., Anselmetti, R., Strait, A., Giulianelli, M. & Ududec, C. Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language. *arXiv preprint arXiv:2507.03409* (2025).

P9. Collins, K. M., Chen, V., Sucholutsky, I., **Kirk**, H. R., Sadek, M., Sargeant, H., Talwalkar, A., Weller, A. & Bhatt, U. Modulating language model experiences through frictions. *arXiv preprint arXiv:2407.12804* (2024).

P10. Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Bartolo, M., **Kirk**, H. R., *et al.* Introducing v0.5 of the AI Safety Benchmark from MLCommons. *arXiv preprint arXiv:2404.12241* (2024).

P11. Derczynski, L., **Kirk**, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R. & Mohammad, S. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190* (2023).

P12. Vidgen, B., Scherrer, N., **Kirk**, H. R., Qian, R., Kannappan, A., Hale, S. A. & Röttger, P. SimpleSafetyTests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370* (2023).

P13. Bailey, H., **Kirk**, H. R. & Howard, P. China's AI Policy: An NLP Approach to Assessing China's Priorities and Governance (2022).

P14. **Kirk**♥, H. Cooperation and Creed: An Experimental Study of Religious Affiliation in Strategic and Societal Interactions (2019).

Reports

R1. Applin, S., Adesso, G., Ashfaq, R., Bai, M., Brammer, M., Fecht, E., Goodman, A., Grossman, S., Groh, M., **Kirk**, H. R., *et al.* GPT-4V(ision) System Card (2023).

R2. Vidgen, B., Chung, Y.-L., Johansson, P., **Kirk**, H. R., Williams, A., Hale, S. A., Margetts, H. Z., Röttger, P. & Sprejer, L. Tracking abuse on Twitter against football players in the 2021–22 Premier League season (2022).

Datasets and Models

D1. Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B. & Hale, S. A. The PRISM Alignment Dataset. *Hugging Face* (2024).
– 40,582 Downloads.

D2. Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T. & Hale, S. A. HatemojiBuild Dataset. *Hugging Face* (2021).
– 1,434 Downloads.

D3. Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T. & Hale, S. A. HatemojiCheck Dataset. *Hugging Face* (2021).
– 984 Downloads.

D4. Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T. & Hale, S. A. Hatemoji Fine-Tuned DeBERTa model. *Hugging Face* (2021).
– 3,367 Downloads.

Selected Media Coverage

2025 *The Atlantic*, Friendship, on Demand
2025 *The Washington Post*, Your chatbot friend might be messing with your mind
2024 *Microsoft*, AI 'for all': How access to new models is advancing academic research, from astronomy to education
2022 *The Guardian*, Seven in 10 Premier League players are sent abusive tweets, study shows
2022 *BBC*, Cristiano Ronaldo & Harry Maguire most abused players on Twitter - report
2022 *The Times*, Ofcom tells social media giants to tackle online abuse of footballers
2022 *Sky News*, Online abuse: New study reveals which footballers are target of most abuse on Twitter
2022 *BBC*, Anti-vax groups use carrot emojis to hide Facebook posts
2022 *The Times*, Carrot emoji helps antivaxers dodge the censors
2022 *WIRED*, DALL-E 2 Creates Incredible Images—and Biased Ones You Don't See
2021 *Sky News*, Antisemitism, racism and white supremacist material in podcasts on Spotify, investigation finds
2021 *Bloomberg Business Week*, Racist Emojis Are the Latest Test for Facebook, Twitter Moderators
2021 *The Telegraph*, Big Tech fails to spot racist posts using emojis, study finds
2021 *Sky News*, Emojis are making it harder for tech giants to track down online abuse
2021 *ITV News*, 'Hatemoji': Racially abusive 'emojis' less likely to be detected online than words, study finds

Policy & Governmental Contributions

2023 **Parliamentary Evidence**, House of Lords Communications and Digital Select Committee
Inquiry: Large Language Models (LLM0074)
Co-authored written evidence with Oxford Internet Institute researchers on LLM development, risks, opportunities, and regulatory approaches

2023 **Ditchley Foundation Conference**, *AI and creative destruction: how will current rapid advances in AI through large 'foundation' models impact on society, the economy and governments?*
Invited to contribute to policy round table alongside representatives from the UK Government, Industry AI Labs and Academia

2022 **Ofcom Consultation Response**, First Phase of Online Safety Regulation

Co-authored response on behalf of The Alan Turing Institute on online safety frameworks, illegal content mitigation, and platform risk assessment approaches

Invited Talks

2026 NYU Seminar, *Longitudinal Human-AI Interaction Research*
2025 NeurIPS EvalEval Workshop
2025 NeurIPS Alignment Tutorial Panel
2024 NeurIPS Workshop on Socially Responsible Language Modelling, *A Tale of Two RCTs: Building a rigorous evidence base on the societal impacts of frontier AI inside the UK Government*
2024 NeurIPS Workshop on Behavioural ML, *Putting the H Back in RLHF: Challenging assumptions of human behaviour for AI alignment*
2024 NeurIPS Workshop on Pluralistic Alignment, *Plugging human data gaps in empirical alignment research*
2024 University of Toronto, Toronto Data Workshop, *The PRISM Alignment Project*
2024 Oxford Generative AI Conference, *How can the UK thrive in the age of generative AI?*
2024 MIT Center for Constructive Communication, *Human-Centric Evaluation of LLMs*
2024 Trinity College, Cambridge, *Challenges of Aligning AI Systems to Human Preferences*
2024 KAIST/NAVER, *What does it mean for AI to be "aligned"?*
2024 Google DeepMind, London, *One Size Fits None? Interrogating Personalised AI Alignment*
2024 Google DeepMind, London, *Plugging human data gaps in empirical alignment research*
2024 Vienna Alignment Workshop, *Plugging human data gaps in empirical alignment research*
2024 AWS, *Plugging human data gaps in empirical alignment research*
2024 EMNLP Custom NLP Workshop, *One Size Fits None? Interrogating Personalised AI Alignment*
2024 Cohere4AI, *A Story of Alignment*
2024 Cohere, *What does it mean for AI to be "aligned"?*
2024 BuzzRobot Community, *What does it mean for AI to be "aligned"?*
2023 OpenAI, *Who decides how LLMs behave? Re-thinking the role of pluralistic human preferences in steering LLMs*
2023 ML2, New York University, *The PRISM Alignment Project*
2023 Royal Statistics Society, Glasgow, *Beneath the hype: How statistics affects language modeling and how language models might affect statisticians in the future*
2022 Oxford University, Connected Life, *Mirror Image: How AI-Generated Images Reflect Our Past and Reshape Our Future*
2022 Bocconi University, *Hate Speech, Bias and Beyond: Who are the humans behind LLMs and why does this matter?*
2022 Manchester United Football Museum, *Tracking online abuse of Premier League footballers with AI*
2021 Truth and Trust Online, *Hatemoji: Understanding and Detecting Online Hate Expressed in Emoji*
2021 AI UK, *The Online Harms Observatory*
2021 GCHQ, *Data-centric AI for online abuse detection*

Academic Advising

2023	Jonathan Ryström, Oxford Internet Institute, University of Oxford, Master's thesis on multilingual and multicultural frontier LLMs (20,000 words) [P6]
2023	Khyati Khandelwal, Oxford Internet Institute, University of Oxford, Master's on casteist biases in frontier LLMs (20,000 words) [C4]
2022	Joint Principal Investigator of the Oxford AI Society Research Labs, producing two top conference papers [W7], [C11]
2023	Joint Principal Investigator of the Oxford AI Society Research Labs, producing two top conference papers [C7], [W6]

Academic Service

Journal Reviewer

Nature; Association of Computational Linguistics; Empirical Methods in Natural Language Processing; NeurIPS; Socially Responsible Language Modelling Workshop (NeurIPS); Workshop on Online Abuse and Harm (ACL*); Nature Machine Intelligence; AI & Society; Data mining and knowledge discovery; Minds and machines

Workshops Organised

2022	Workshop on Online Abuse and Harms (NAACL)
2022	Workshop on Dynamic and Adversarial Data Collection (NAACL)

References Available Upon Request

Last updated: May 11, 2026