XAVIER ROBERTS-GAAL

Harvard Department of Psychology, William James Hall
33 Kirkland Street, Cambridge, MA, USA 02138
xavierrobertsgaal@g.harvard.edu | scholar.harvard.edu/xrg

EDUCATION

Harvard University **2022 – present**
PhD, Psychology (in progress). AM, Psychology (May 2025). Supervisor: Fiery Cushman (link).

- Mentorship: Mira Becker (Honors Thesis, received Psychology Faculty Prize; now at OCD Institute), Nikola Jurkovic (Special Concentration in AI and Society; now at Model Evaluation and Threat Research, the premier AI evaluations nonprofit), Marija Bolic (UOttawa, summer internship; now PhD student at University of Toronto)

University of Oxford (New College), Oxford, UK **2017 – 2020**
BA (Honours) Psychology & Philosophy, First Class Honours

- Thesis: 'Two rival pictures of social cognition'. Supervisor: Anita Avramides
- Key classes: How to Build a Brain (Neuroscience + AI), Clinical Social Cognitive Neuroscience
- Activities: Oxford Nightline (crisis hotline), Oxford Union (debating prize), US debate coaching

AWARDS AND HONORS

- National Defense Science and Engineering Graduate Fellowship, US Air Force Office for Scientific Research, >$200,000, *accepted*
- National Science Foundation Graduate Research Fellowship, >$150,000, *declined*
- Lightspeed (Jaan Tallinn), $15,000, awarded for project on dual-process cognition for AI safety
- Norman Henry Anderson Graduate Psychology Fund at Harvard, $7,500, awarded for cross-cultural research project with the Himba in Kunene Region, Namibia
- Stimson Fund at Harvard, $1,000, awarded for projects in moral decision-making
- Harvard Mind Brain Behavior Interfaculty Initiative, Graduate Student Award, $820, for "Compositionality in cognition: the language of thought and its discontents"
- AI Safety Student Team at Harvard, $300, awarded for project on measuring cognitive capacities in LLM agents
- New College, Oxford Examination Prize for scholarship in Finals
- USA National Merit Scholar Finalist
- NCFL Grand National Tournament, 2x Semifinalist (top 3 high school Lincoln-Douglas debater in the USA at national championship)

PAPERS [*shared first-authorship. Tags: ^{m} (moral cognition), ^{c} (culture and social cognition), ^{AI} (AI)]

Working papers

1. $^{AI, m, c}$ Voudouris, K*, **Roberts-Gaal, X*,** Buhl, M., Irving, G., & Summerfield, C. AI alignment is a human problem. https://osf.io/preprints/psyarxiv/zqngj_v1
2. ^{m} Le Pargneux, A.*, **Roberts-Gaal, X.***, & Cushman, F. A. Hard bargains and even splits: Fairness judgments track bargaining power across diverse cultures. Under invited revision at the *Proceedings of the National Academy of Sciences*. https://osf.io/preprints/psyarxiv/3uqks_v2

3. [AI] Zhou Hong, S., Kleinman, A., Mathiowetz, A., Howes, A., Cohen, J., Ganta, S., Letizia, A., Liao, D., Pahari, D., **Roberts-Gaal, X.**, Righetti, L., & Torres, J. (2026). Measuring mid-2025 LLM-assistance on novice performance in biology. https://arxiv.org/abs/2602.16703 [first large-scale experiment of its kind; papers in AI evaluation are often published only on ArXiv]
4. [m] Law, K. F., …, **Roberts-Gaal, X.,** …, & Syropoulos, S. (2026). Social discounting outpredicts temporal discounting in intergenerational judgment: A many labs investigation. https://osf.io/6qyv7_v1
5. [m, c] **Roberts-Gaal, X.*,** Le Pargneux, A.*, Mungunda, M., Tjikuvua, P., Fredrik, C., Hartley, V., Cushman, F. A., Kroupin, I., & Davis, H. Moral judgment and proportional contribution in a small-scale society.
6. [AI] **Roberts-Gaal, X.**, Williams, T., Turner, J. M., Mykhailiuk, J., Wisdom, S., Duus, E., Wei, K., & Kolt., N. Managing multi-agent AI risks: Lessons from energy infrastructure and financial markets.
7. [c] **Roberts-Gaal, X.*,** Zeng, T. C.*, Mungunda, M., Tjikuvua, P., Fredrik, C., Hartley, V., Henrich, J., Cushman, F. A., Kroupin, I., & Davis, H. Model-free and model-based control beyond WEIRD.
8. [AI] Paskov, P.*, Wei, K.*, Hong, S. Z., Bateyko, D., Ezell, C., **Roberts-Gaal, X.,** Guest, E. M., & Bhatt, E. RCTs & human uplift studies: Methodological challenges and practical solutions for frontier AI evaluation. Under review at *AI Ethics and Society* 2026. https://arxiv.org/abs/2603.11001

Published papers

9. [c] **Roberts-Gaal, X.**, Bolic, M., & Cushman, F.A. (2025). Environmental variability shapes the representational format of cultural learning. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.250528312
10. Woodley, L.*, **Roberts-Gaal, X.***, Calcott, R.*, & Cushman, F. A. (2026). Experimenter demand effects in online psychology experiments. In press at *Open Mind*.
11. [AI] Ezell, C., **Roberts-Gaal, X.**, & Chan, A. (2025). Incident Analysis for AI Agents. *Proceedings of the AAAI ACM Conference on AI, Ethics, and Society*, *8*(1), 865–878. https://doi.org/10.1609/aies.v8i1.36596
12. [AI] Wei, K.*, Paskov, P.*, Dev, S.*, Byun, M.*, Reuel, A., **Roberts-Gaal, X.**, Calcott, R., Coxon, E., & Deshpande, C. (2025). Position: Human baselines in model evalutaions need rigor and transparency (with recommendations & reporting checklist). *International Conference on Machine Learning* (spotlight poster). https://github.com/kevinlwei/human-baselines
13. [c] Levy, A., **Roberts-Gaal, X.,** & Cushman, F. A. (2025). No Evidence for Cost-Benefit Arbitration Between Social Learning Strategies. *Proceedings of the Cognitive Science Society, 47.* https://escholarship.org/uc/item/6gs0v9rs
14. [c] **Roberts-Gaal, X.** & Cushman, F. A. (2023). Computational Principles Underlying the Evolution of Cultural Learning Mechanisms. *Proceedings of the Cognitive Science Society, 45.* https://escholarship.org/uc/item/78c4w10j
15. Cuve, H. C., Stojanov, J., **Roberts-Gaal, X.**, Catmur, C., & Bird, G. (2021). Validation of Gazepoint Low-Cost Eye-Tracking and Psychophysiology Bundle. *Behavior Research Methods, 54,* 1027–1049. https://doi.org/10.3758/s13428-021-01654-x

CONFERENCE PRESENTATIONS (first-author only)

- "No evidence for experimenter demand effects in three online psychology experiments." *Society for Philosophy and Psychology,* Johns Hopkins University, June 2026 (Poster)
- "A theory of cultural simplification." *Society for Philosophy and Psychology,* Johns Hopkins University, June 2026 (Poster)
- "Hard bargains and even splits: Fairness judgments track bargaining power across diverse cultures." *Human Behavior and Evolution Society*, **finalist for Young Investigator award**, Rabat, Morocco, May 2026 (Talk)
- "A theory of cultural simplification." *Cultural Evolution Society*, Rabat, Morocco, May 2026 (Talk)
- "Third-party moral judgments about negotiations are sensitive to bargaining power." *Cognitive Science Society*, San Francisco, USA, July 2025 (Poster)
- "Third-party moral judgments about negotiations are sensitive to bargaining power." *Society for Philosophy and Psychology*, Carnegie Mellon University, June 2025 (Poster)
- "Environmental variability shapes the cognitive engine of culture." *Society for Philosophy and Psychology*, Purdue University, June 2024 (Poster)
- "Computational principles of cultural learning." *Cognitive Science Society*, Sydney, Australia, July 2023 (Talk)
- "Uncertainty in moral judgment." *Society for Philosophy and Psychology*, University of Pittsburgh, June 2023 (Poster)
- "Uncertainty in moral judgment." *European Experimental Philosophy Conference*, University of Zürich, August 2023 (Poster)

PROFESSIONAL EXPERIENCE

Advisor, confidential clients **Nov 2025 – Present**

- *Ad hoc* grantmaking advisory, including portfolio evaluation and impact strategy. Largely in biosecurity and AI resilience.

Prof. Zoë Johnson King **Sep 2023 – May 2024**
Research Assistant, Philosophy Department

- Research assistance, including distilling the relevant psychology/neuroscience literature, for Prof. Johnson King's book, *Praiseworthiness: Its Contours and Limits*

McKinsey & Company, London, UK; Global projects **Nov 2020 – Jun 2022**
Business Analyst

- Helped manage creation and deployment of 4 machine learning models driving £200m revenue impact for top UK bank. Won key stakeholder support. Identified, designed, and implemented 3 marketing initiatives to capture high potential customer segments
- Defined technical architecture, governance, and data ethics approach for cross-govt data linking effort (50+ datasets) joining up social services to support vulnerable individuals
- Created P&L model and identified ~£50m store productivity uplift for UK retailer's strategy
- Designed and modelled incentives to boost efficacy of £200m spend for top-5 auto group
- Conducted 3 due diligences, on packaging, ecommerce/fashion, and translation companies

- Led training sessions for 10k employees in largest Agile transformation in the energy sector
- Consistently ranked in top quintile of performance. Contributed to 3 proposal wins

Summer Business Analyst **Jun 2019 – Aug 2019**

- Implemented analytics system in top-10 EU corporate bank to improve capital management
- Conducted growth strategy and due diligence for heritage footwear company which later IPO'd

Human Information Processing and Neurotheory Labs, Oxford, UK (link) **Jun 2020 – Nov 2020**
Research Assistant – PIs: Chris Summerfield, Andrew Saxe

- Investigated the learning dynamics of deep neural networks parameterized by Hebbian and top-down feedback components to identify which rules yield qualitative signatures of humanlike learning (e.g., stagelike transitions in loss function, illusory correlations)
- Used formal analysis, representational similarity analysis, and multidimensional scaling, and developed pipelines in Python (JAX, TensorFlow)
- Scaled up experiment using texture synthesis patterns and naturalistic computer graphics
- Worked on project implementing meta-learning of intrinsic reward to model mood disorders
- Contributed to journal club and lab activities

Bird Lab, Oxford, UK (link) **Apr 2019 – Jun 2020**
Research Assistant – PI: Geoff Bird

- Co-authored a new validation protocol for low-cost eye-trackers used in psychophysics (Gazepoint GP3HD and biometric package). Identified ways to minimize data loss
- Gathered data for experiments on interoception, emotional gaze in autism, and empathy
- Built an R-Shiny web app to process and analyze data, improving the lab's workflow efficiency

Moonshot Global Consulting, Washington, DC, USA; Global projects **Jan 2018 – Jun 2019**
Intern

- Supported on monitoring and evaluation projects for Australian, USA, and UK development agencies, incl. AUD 15 million project growing the APAC impact investing ecosystem
- $400,000 World Bank project implementing a mechanism for private sector feedback on public services, and $45 million, 4.5-year open innovation project funded by USAID, DfID, Sida, and Omidyar Network

SERVICE

- *Ad hoc* peer review for *Nature Communications, Journal of Personality and Social Psychology*, *Evolution & Human Behavior, Synthese, Cognitive Science Society*, *AI Ethics and Society*, *NeurIPS MP2 Workshop*
- Facilitator, **AI Safety Student Team**, AI safety fundamentals for Harvard/MIT undergrads
- Mentor, **PPREP** (supporting PhD/RA applicants from underrepresented groups)
- Convened interdisciplinary PhD reading group on **compositionality** in cognition, Spring 2023
- Created professional PowerPoint templates for lab

SKILLS

Analytical	Statistical modeling (incl. Bayesian methods), deep learning, computational cognitive models, evolutionary models, game theory, RL, LLMs
Languages (programming)	Python (PyTorch, TensorFlow, JAX, NumPy, Pandas, Scikit-Learn, PsychoPy, PyGaze), R, Julia, intermediate MATLAB
Software	Tableau, SPSS, JASP, Alteryx, Word, Excel, PowerPoint, Qualtrics, Gorilla, Git, Atlassian tools (e.g., Jira, Confluence, Trello, Bitbucket), Azure DevOps
Hardware	Gazepoint GP3HD eye-tracker and biometrics bundle, Biopac
Languages (human)	Native English, intermediate Mandarin
Interpersonal	Leadership, project management, empathic/crisis listening, senior executive communication, coaching, workshop facilitation, and interviewing

ADDITIONAL INFORMATION

Citizenship	USA and Australia citizenship
Interests	Rock climbing, rowing, travel, swimming, poetry, piano, tea, reading, and thinking a lot about minds
Volunteering	Co-organizer, UK AI Security Institute Alignment conference; advised Taimaka, a Nigeria-based hunger/poverty charity (recipient of D-Prize, funded by USAID); mentored neuroethics debates for high schoolers