Joan Velja

AI alignment · scalable oversight · adversarial elicitation

joan.velja22@gmail.com
joanvelja.com
Google Scholar
GitHub

Education

2025–2028 D.Phil. in Computer Science, University of Oxford
Advised by Alessandro Abate (OxCAV Group).

2023–2025 M.Sc. Artificial Intelligence, *cum laude* (GPA: 8.6/10), University of Amsterdam

2023 B.Sc. Artificial Intelligence (GPA: 4.0/4.0), University of Technology Sydney
Exchange semester. Merit scholarship (full ride).

2020–2023 B.Sc. Data Science (110/110 cum laude), Università Bocconi
Thesis: *The Double Descent Phenomenon in Neural Networks*. Supervisor: Enrico Maria Malatesta.

Experience

2026– PI, UK AISI Alignment Project, University of Oxford
Adversarial Elicitation Theory. Funded by the Alignment Project (£300k).

2026– Scholar, ML Alignment & Theory Scholars (MATS), Berkeley
Scalable Oversight (Debate). Mentored by Jacob Pfau (UK AISI) and Shi Feng (GWU).

2025–2026 Project co-supervisor, LASR Labs, London
Red-teaming Untrusted Monitoring. Co-supervision with Charlie Griffin (UK AISI).

2025 Thesis, MSc Honors Programme — University of Oxford
Prover-Verifier Games for AI Control. Supervisor: Alessandro Abate.

2024 AI Alignment Researcher, LASR Labs, London
Steganographic collusion in multi-agent AI systems. Supervisors: Nandi Schoots, Dylan Cope, Charlie Griffin.

Selected Publications

2025 When Can We Trust an Untrusted Monitor?
Gardner-Challis N.*, Bostock J.*, Kozhevnikov G.*, Sinclaire M.*, Velja J., Griffin C., Abate A. *Preprint*

2025 Prover-Verifier Games for AI Control
Velja J.*, Griffin C., Abate A. *Preprint*

2025 Modular Training of Neural Networks aids Interpretability
Golechha S.*, Chaudhary M.*, Velja J.*, Abate A., Schoots N. *Under Review*

2025 Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs
Mathew Y.*, Matthews O.*, McCarthy R.*, Velja J.*, Schroeder de Witt C., Cope D., Schoots N. *AACL 2025 (Oral), NeurIPS 2024 Workshop*

2024 Dynamic Vocabulary Pruning in Early-Exit LLMs
Velja J.*, Abdel Sadek K.*, Nulli M.*, Vincenti J.*, Jazbec M. *ENLSP-IV NeurIPS Workshop 2024*

2024 Explaining RL Decisions with Trajectories: A Reproducibility Study
Velja J.*, Abdel Sadek K.*, Nulli M.*, Vincenti J.*. *TMLR 2024*

2024 Assessing Expertise Overlap in Mixture of Experts Architectures
Velja J.*, Divak A.*. *MechInterp Hackathon*

Awards & Grants

2026 Research Grant, UK AISI Alignment Project, Adversarial Elicitation Theory, £300,000
2025 Research Grant, Open Philanthropy, Graduate Studies funding, $380,000
2025 Research Grant, Open Philanthropy, Prover-Verifier Games for AI Control, $25,000
2024 Fellowship, Arcadia Impact, LASR Labs AI Safety Research Fellowship, $15,000
2024 Travel Grant, Human Aligned AI Summer School, $1,000
2020–2022 Merit Scholarship, Università Bocconi, *top 2%*, 20,000€
2022 3rd Place, Brembo Sensify Hackathon, 2,000€

Talks

2025 *Building an Empirical Science of AI Debate*, MATS Research Symposium, Berkeley (Spotlight)
2025 *Why Should We Care About Steganography?*, ControlConf 2025

Reviewing

2026 ICML 2026
2025 IASEAI 2025 (International Association for Safe and Ethical AI)

Projects

2024 Pitfalls from Partial Observability in RLHF
Empirical analysis of deceptive inflation and overjustification under partial observability. Supervision: Leon Lang (UvA), Davis Foote (CHAI). *[Report]*

Extracurriculars

2022 Co-founder, Bocconi AI & Neuroscience Student Association
2021–2023 Lead Manager, Think Tank Tortuga
2016–2019 Team Captain, Rugby Alto Vicentino

Technical Skills

Programming: Python, C++, R, Julia, C#, SQL, Stata
Frameworks: PyTorch, TensorFlow, TransformerLens, Transformers, TRL, Gym, StableBaselines, d3rlpy
Languages: Albanian, Italian (Native); English (Fluent); German (Intermediate)