Joan Velja

☎ +39 3319261802 ✉ joan.velja22@gmail.com in linkedin.com/joanvelja github.com/joanvelja
joanvelja.com 🎓 scholar.google.com/joanvelja

Education

University of Oxford Exp. 2028
D.Phil. in Computer Science *Oxford, United Kingdom*

- Incoming D.Phil. student, starting in Fall 2025. Working on AI Alignment and Elicitation.
- Supervisor: **Alessandro Abate**, OxCAV Group.

University of Amsterdam July 2025
M.Sc. Artificial Intelligence (GPA: 8.6/10, 4.0 equivalent. Graduated Cum Laude) *Amsterdam, Netherlands*

- **Relevant Coursework:** Machine Learning, Natural Language Processing I, Deep Learning I-II, Foundation Models, Reinforcement Learning
- **Research interests:** AI Alignment, Reinforcement Learning

University of Technology Sydney (Exchange Semester) June 2023
B.Sc Artificial Intelligence (GPA: 4.0/4.0) *Sydney, Australia*

- Selected by academic merit to participate in the Exchange Program 2022-2023 with a full ride scholarship.
- **Relevant Coursework:** Deep Learning and Convolutional Neural Network, Natural Language Processing, Cloud computing and software as a service, Data Structures and Algorithms

Università Bocconi July 2023
B.Sc Data Science (GPA: 29.3/30, 4.0 equivalent); Graduated 110/110 Cum Laude *Milan, Italy*

- **Relevant Coursework:** Fundamentals of Computer Science, Computer Programming, Machine Learning, Mathematics, Advanced Mathematics, Statistics, Advanced Statistics, Big Data and Databases, Game Theory
- **Awards:** Awarded for 3 years scholarship for deserving students, top 2% in my cohort.
- **Thesis:** An unorthodox shift in the variance-bias tradeoff in Neural Networks: the double descent phenomenon and the ease of training in the overparametrized regime, grade: 4/4. Read about it *here*.
- Thesis Supervisor: **Enrico Maria Malatesta**

Experience

LASR Labs Aug 2025 – Present
Project co-supervisor *London, United Kingdom*

- **Topic:** Red-teaming Untrusted Monitoring.
- Developing a more accurate control evaluation of untrusted monitoring, particularly by relaxing the assumption that a human red-team can provide a coordination strategy.
- Co-supervision with **Charlie Griffin** (UK AISI); providing technical and experimental guidance.

MSc Honors Programme - University of Oxford Jan 2025 – Jul 2025
Thesis *Oxford, United Kingdom*

- **Project topic:** Prover-Verifier Games for AI Control.
- Introducing sequentiality in prover play – as opposed to simultaneous play in Kirchner et al. (2024) – as a measure for mitigating collusion/coordination between players and avoid equilibria caused by spurious correlations.
- Studying generalization of Verifier capabilities to arbitrary strong provers out-of-distribution.

- Supervisor: **Alessandro Abate**.

LASR Labs Jul 2024 – Oct 2024
AI Alignment Researcher *London, United Kingdom*

- 12 weeks program aimed at producing a paper and accompanying blog post that makes incremental progress on an important problem in AI safety.
- Project revolved around elicitation of *steganography* in multi-agent AI systems under optimization pressure (In Context Learning , Reinforcement Learning) drawing from the AI Control research agenda proposed by Redwood Research and Model Organisms by Anthropic.
- Assessed whether steganography could emerge as an instrumental goal and to red-team paraphrasing, exploring alternative mitigation strategies.
- First paper to demonstrate instrumental emergence of steganographic collusion as a means to achieve a goal.
- Supervisors: **Nandi Schoots**, **Dylan Cope**, **Charlie Griffin**.
- Find our preprint *clicking here* and slides to presentation *clicking here*.

1000 Kelvin GmbH Aug 2023 – Dec 2023
R&D Software Engineer *Berlin, Germany*

- Developed a Large Language Model (Llama-2 Finetune) specialized in Physics and Additive Manufacturing.
- Improved the company's proprietary model training pipeline, through a new approach towards generation and utilization of synthetic data.

World Food Programme Jun 2022 – Sep 2022
Data Scientist - Intern *Rome, Italy*

- Developed analytical dashboards in Tableau, still in use in the Head of Business Continuity department, by leveraging unused organizational data.
- Applied predictive classification algorithms to dashboard data, resulting in improved accuracy in predicting country adequacy levels, to enhance organizational strategic planning.
- Conducted a comprehensive comparative analysis of countries in North West Africa from a Business Continuity perspective, leading to identification of key operational gaps or best practices, which informed the fine-tuning of dashboard use cases and was presented to the heads of the Regional Bureau in Dakar, Senegal.

Projects

Pitfalls from Partial Observability in RLHF | *Reinforcement Learning, Stable Baselines 3*

- In recent work, a theoretical analysis of the consequences of partial observability in RLHF has been made: agents may deceptively inflate performance, or overjustify behavior to make an impression. In this project, I repeated the analysis of the paper by Lang et. al (2024), but focusing on empirical demonstrations of the failure modes and solutions, thus complementing the existing theoretical analysis. I've done so by implementing a synthetic human in toy tasks to make "preference choices" based on partial observations, by e.g. modeling the human as Boltzmann rational. Moreover, I conducted variations on partial observability frameworks, with different experimental setups.
- Find the output of this project *here*.
- Supervision by: **Leon Lang** (University of Amsterdam), **Davis Foote** (Center for Human-Compatible AI, UC Berkeley).

Dynamic Vocabulary Pruning in Early-Exiting LLMs | *Large Language Models, Pytorch*

- This project explores the effectiveness of early exiting strategies in Large Language Models (LLMs) during inference to reduce computational overhead. By allowing the model to exit at specific points in the attention stack based on confidence thresholds, we seek to maintain performance while reducing runtime. The core idea of the project is to speed up the *unembedding* operation during inference by projecting onto a subset of dimensions. This is of interest as a natural advancement to the original work of Schuster et al. 2022. The project, accepted as a workshop paper at ENLSP-IV NeurIPS 2024, shows how we can attain significant gains in FLOPs and latency, while allowing the model to exit several significantly earlier and retaining up to 100% of its performance.
- Find the paper – accepted to ENLSP-IV NeurIPS 2024 – *here*.
- Supervision by: **Metod Jazbec**.

Publications

Gardner-Challis N.*, Bostock J.*, Kozhevnikov G.*, Sinclaire M.*, **Velja Joan**, Griffin C., Abate A. *'When can we trust an untrusted monitor?', Preprint*

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

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

Mathew Y.*, Matthews O.*, McCarthy R.*, **Velja Joan***, Schroeder de Witt C., Cope D., Schoots N. *'Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs', AACL 2025, Towards Safe & Trustworthy Agents, NeurIPS 2024 Workshop*

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

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

Velja Joan*, Divak Adam* *'Assessing expertise overlap in Mixture of Experts Architectures', MechInterp Hackathon Project*

Awards, Honors and Scholarships

Research Grant | *Open Philanthropy, Graduate Studies funding, 380,000 $*	2025
Research Grant | *Open Philanthropy, Prover-Verifier Games for AI Control, 25,000 $*	2025
Fellowship | *Arcadia Impact, LASR Labs: AI Safety Research Fellowship Summer '24, 15,000 $*	2024
Scholarship and Travel Grant | *Human Aligned AI Summer School, 1,000 $*	2024
Scholarship | *Università Bocconi: Merit Scholarship, 20,000€*	2020, 2021, 2022
Award | *Brembo Sensify Hackathon, 3rd Place, 2,000€*	2022
Scholarship | *Comune di Schio: Merit Scholarship, top 5% of cohort for the class of 2019, 1,000€*	2019
Scholarship | *Banca Intesa San Paolo: Merit Scholarship for outstanding academic results, 500€*	2015, 2016, 2017, 2018
Scholarship | *Regione Veneto: Merit Scholarship for outstanding academic results, 300€*	2014

EXTRACURRICULARS

Co-founder | *Bocconi AI & Neuroscience Student Association* 2022

Team Formation, Lead Manager | *Think Tank Tortuga* 2021-23

Team Captain | *Rugby Alto Vicentino* 2016-19

TECHNICAL SKILLS

Programming Languages: Python, C++, R, Julia (proficient) - C#, HTML, SQL, Stata, Tableau (advanced) - CSS (familiar)
Frameworks: Vue.js, Django, TensorFlow, PyTorch, TransformerLens, Gym, StableBaselines, d3rlpy, Transformers, TRL
Languages: Albanian, Italian (Native), English (Fluent), German (Intermediate)

LINKS

Bachelor's Thesis: https://joanvelja.com/static/pdf/BSc_Thesis.pdf
Hidden in Plain Text (Steganography paper): https://arxiv.org/abs/2410.03768
PORLHF : https://joanvelja.com/static/pdf/PORLHF.pdf
Dynamic Pruning : https://arxiv.org/abs/2410.18952