

# JOAN VELJA

📞 +39 3319261802    ✉️ joan.velja22@gmail.com    🔗 linkedin.com/joanvelja    🐙 github.com/joanvelja  
🌐 joanvelja.com    🎓 scholar.google.com/joanvelja

## EDUCATION

---

### University of Oxford

*D.Phil. in Computer Science*

Exp. 2028

*Oxford, United Kingdom*

- D.Phil. (PhD) student. Working on AI Alignment and Elicitation.
- Supervisor: **Alessandro Abate**, OxCAV Group.

### University of Amsterdam

*M.Sc. Artificial Intelligence (GPA: 8.6/10, 4.0 equivalent. Graduated Cum Laude)*

July 2025

*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)

*B.Sc Artificial Intelligence (GPA: 4.0/4.0)*

June 2023

*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

*B.Sc Data Science (GPA: 29.3/30, 4.0 equivalent); Graduated 110/110 Cum Laude*

July 2023

*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**

## PUBLICATIONS

---

Gardner-Challis N.\*, Bostock J.\*, Kozhevnikov G.\*, Sinclair 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 (Oral), 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

## EXPERIENCE

---

### ML Alignment & Theory Scholars

*Scholar*

Jan 2026 – Present

*Berkeley, California*

- **Project Topic:** Scalable Oversight (Debate) and Generalization (Sycophancy training).
- Mentored by **Jacob Pfau** (UK AISI) and **Shi Feng** (George Washington University).

### LASR Labs

*Project co-supervisor*

Aug 2025 – Jan 2026

*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

*Thesis*

Jan 2025 – Jul 2025

*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

*AI Alignment Researcher*

Jul 2024 – Oct 2024

*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*.

## PROJECTS

---

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

- Empirical analysis of partial observability in RLHF: agents may deceptively inflate performance or overjustify behavior. Implemented synthetic human evaluators in toy tasks using Boltzmann rationality.
- Supervision: **Leon Lang** (UvA), **Davis Foote** (CHAI, UC Berkeley). *[Report]*

## AWARDS, HONORS AND SCHOLARSHIPS

---

<b>Research Grant</b>   <i>Open Philanthropy, Graduate Studies funding, \$380,000</i>	2025
<b>Research Grant</b>   <i>Open Philanthropy, Prover-Verifier Games for AI Control, \$25,000</i>	2025
<b>Fellowship</b>   <i>Arcadia Impact, LASR Labs AI Safety Research Fellowship, \$15,000</i>	2024
<b>Scholarship</b>   <i>Human Aligned AI Summer School, travel grant, \$1,000</i>	2024
<b>Scholarship</b>   <i>Università Bocconi, Merit Scholarship, 20,000€</i>	2020–2022
<b>Award</b>   <i>Brembo Sensify Hackathon, 3rd Place, 2,000€</i>	2022
<b>Scholarship</b>   <i>Comune di Schio, Merit Scholarship, top 5%, 1,000€</i>	2019
<b>Scholarship</b>   <i>Banca Intesa San Paolo, academic merit, 500€</i>	2015–2018
<b>Scholarship</b>   <i>Regione Veneto, academic merit, 300€</i>	2014

## EXTRACURRICULARS

---

<b>Co-founder</b>   <i>Bocconi AI &amp; Neuroscience Student Association</i>	2022
<b>Lead Manager</b>   <i>Think Tank Tortuga</i>	2021–2023
<b>Team Captain</b>   <i>Rugby Alto Vicentino</i>	2016–2019

## TECHNICAL SKILLS

---

**Programming Languages:** Python, C++, R, Julia (proficient); C#, HTML, SQL, Stata, Tableau (advanced); CSS (familiar)

**Frameworks:** PyTorch, TensorFlow, TransformerLens, Transformers, TRL, Gym, StableBaselines, d3rlpy, Vue.js, Django

**Languages:** Albanian, Italian (Native); English (Fluent); German (Intermediate)