Dr. Katarina Slama

PhD Neuroscience, University of California, Berkeley

LinkedIn · Google Scholar

AI Research Scientist

Experienced AI Research Scientist (PhD) specializing in empirical AI safety research: model evaluations, alignment, and frontier model behavior. I helped build InstructGPT at OpenAI and led cross-functional, preregistered model-behavior research at the UK AI Security Institute. My current work is focused on the psychometric properties of loss of control risk factors. My publications are widely cited and I am regularly invited to speak internationally on AI safety.

Professional Experience

Research Scientist, AI
Rethink Priorities UK, seconded to UK AISI *2025 – 2026*

- Conducted empirical research on model behavior, including preference-driven model action.
- Designed and executed scientifically rigorous model evaluations, including preregistered studies.
- Led cross-functional research collaborations to deliver actionable results across AI subdomains.
- Delivered invited research talks internationally, including the opening keynote at the Open Data Science Conference (San Francisco) and AI safety events in Helsinki and Paris.

Member of Technical Staff on Alignment and Policy Research
OpenAI, Inc. *2021 – 2024*

- Analyzed human labeler data to improve the InstructGPT models (precursors to ChatGPT).
- Co-authored the InstructGPT paper, including the model card covering societal impacts.
- Created evaluation frameworks and conducted in-depth analyses of model behavior to improve reliability and user outcomes.
- Owned the section on public figure generation for the DALL·E 3 system card.

Data Scientist on Security, Risk and Fraud
Intuit, Inc. *2020 – 2021*

- Significantly improved detection of risk behaviors, delivering an 18-fold improvement in accuracy.
- Led cross-functional, international team meetings with data scientists, software engineers, and risk specialists, ensuring smooth collaboration and project progression.
- Developed, maintained, and deployed machine learning models, including language models, on massive datasets using Python, Spark, and AWS.

Researcher in Neuroscience
UC Berkeley *2020*

- Published first-author paper on the neural mechanisms of visual search in humans.
- Coauthored paper on gender bias in academia, proposing actionable solutions.

Scholar in AI and Neuroscience
OpenAI, Inc. *2020*

- Completed competitive deep learning training program (<2% admission rate).
- Developed a deep learning project for epileptic seizure prediction using neural networks.

Laboratory Manager in Psychology
Harvard University *2011 – 2013*

- Facilitated research projects in MacArthur award-winning laboratory.
- Contributed to the end-to-end research life cycle, including: idea generation, grant applications, experimental design, data collection, statistical analysis, paper writing, reviews and publication.
- Helped supervise approximately 30 research assistants, including interviewing and allocating to best-fit projects, as well as continuous mentoring and support.

Skills

Programming: Python (numpy, pandas, matplotlib, scikit-learn, pytorch, tensorflow), Git, Matlab
Data Science: Machine Learning, Deep Learning, NLP, Statistical Analysis, Data Visualization
Languages: Swedish (native), English (fluent), Finnish (proficient), German (proficient)

Education

University of California, Berkeley
Ph.D. in Neuroscience, Helen Wills Neuroscience Institute *2013-2019*
Dissertation: Mechanisms of visual attention, and their relationships to expectation and navigation, in the human brain

Brown University, Providence, RI
Bachelor of Science in Psychology *2007-2011*
Magna Cum Laude, Honors in Concentration

Selected Publications

- **Slama, K.**, Souly, A., Bansal, D., Davidson, H., Summerfield, C., & Luettgau, L. (2026). *When Do LLM Preferences Predict Downstream Behavior?*. arXiv preprint arXiv:2602.18971. ↪
- 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. (2025). *Lessons from a chimp: AI "scheming" and the quest for ape language*. arXiv preprint arXiv:2507.03409. ↪
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., **Slama, K.**, Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *NeurIPS*. ↪
- **Slama, S. J. K.**, Jimenez, R., Saha, S., King-Stephens, D., Laxer, K. D., Weber, P. B., Endestad, T., Ivanovic, J., Larsson, P. G., Solbakk, A.-K., Lin, J. J., & Knight, R. T. (2021). Intracranial Recordings Demonstrate Both Cortical and Medial Temporal Lobe Engagement in Visual Search in Humans. *Journal of Cognitive Neuroscience*, *33*(9), 1833–1861. ↪