Felix Jedidja Binder

me@felixbinder.net ac.felixbinder.net +1 (858) 291-2056 Full CV

Education

2019–2025 **University of California San Diego** PhD in Cognitive Science
2024-2025 **Stanford University** Visiting Researcher
2013–2019 **Freie Universität Berlin** Bachelor of Arts in Philosophy & Computer Science

Experience

2025 Menlo Park **Research Scientist** | Meta | *Meta Superintelligence Labs*

- AI Safety & Alignment for TBD Labs

2025 San Francisco **Research Scientist** | Scale AI | *Safety, Evaluations & Alignment Lab*

- Leading development of benchmark measuring active value learning in LLMs for public evaluation.
- Contributing to economic impact evaluation of Computer Use Agents.

2019–2024 San Diego **Graduate Student Researcher** | University of California San Diego | *Cognitive Science Department*

- Created and maintained a full stack setup for running web experiments evaluating human and AI behavior on a range of cognitive tasks (Cognitive AI Benchmarking).
- Led a study comparing humans and planning algorithms on a simulated physical construction task.
- Created a dataset for a large benchmarking study of physical understanding in humans & AI (Physion) with NeuroAILab (Stanford) and Computational Cognitive Science lab (MIT).
- Evaluated a broad suite of state-of-the-art vision & particle-based AI models on the Physion dataset. Found that AI models do not yet meet human performance in physical understanding.
- Created public outreach videos on neural networks and AI ethics for high school students with pathways2AI.
- Taught undergrad & graduate courses, including *Reinforcement Learning* and *Data Science*.
- Organized the Cognitive AI Benchmarking workshop at the 45th Annual Meeting of the Cognitive Science Society.

2024 Berkeley **AI Safety Research Fellow** with Owain Evans | Constellation Astra Fellowship

- Developed novel experimental framework to train and evaluate introspection in large language models (LLMs).
- Demonstrated that frontier LLMs (GPT-4, GPT-4o, Llama 3 70B) can acquire knowledge about themselves through introspection, not just from training data.

2023 Cambridge, MA **AI Research Scientist Intern** | Cambria Labs

- Oversaw creation of multimodal video dataset for physical understanding and prediction.
- Built a data pipeline for data management & model training; implemented and trained a suite of vision transformer based models on the dataset.

2023 **Artificial General Intelligence Safety Fundamentals Course** | BlueDot Impact

- Developed an evaluation protocol that isolates causal effects of context for analyzing steganographic tendencies (covert information encoding) in large language models.
- Conducted an investigation into potential steganographic behavior in current LLMs, utilizing the aforementioned evaluation protocol.

Skills

Programming & AI Python & PyTorch, AI Safety & Alignment, RL, LLMs, Interpretability, Planning & Reasoning, Evals
Statistics Experiment Design, Model Fitting & Analysis, Hypothesis Testing, Bayesian Statistics
Communication Scientific Writing, Public Science Communication, Data Visualization, Cross-Field Communication

Selected Publications

* indicates equal contribution.

2025 **Binder, F.**, Mattar, M., Kirsh, D., & Fan, J. Humans Select Subgoals That Balance Immediate and Future Cognitive Costs During Physical Assembly. *Cognitive Science.*

2025 **Binder, F.** Thinking Through Action: Prediction, Planning, and Metacognition in Problem-Solving. *Doctoral dissertation, University of California, San Diego.* Dissertation

2024 **Binder, F.***, Chua, J.*, Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. Looking Inward: Language Models Can Learn About Themselves by Introspection. *ICLR 2025*. | Code & Paper

2021 Bear, D.*, Wang, E.*, Mrowca, D.*, **Binder, F.***, Tung, H., Pramod, R. T., Holdaway, C., Tao, S., Smith, K., Sun, F., Fei-Fei, L., Kanwisher, N., Tenenbaum, J., Yamins, D.** & Fan, J.** Physion: Evaluating Physical Prediction from Vision in Humans and Machines. *NeurIPS 2021 (Datasets & Benchmarks track)* | Code & Paper, NeurIPS Presentation