Felix Jedidja Binder

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

Education

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

Experience

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

Experiment Design

- Created and maintained a full stack setup for running web experiments comparing human and AI behavior on a range of cognitive tasks (Cognitive AI Benchmarking).
- Designed, implemented and conducted a web-based study to compare humans and planning algorithms on a simulated physical construction task.

Artificial Intelligence

- 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.

Teaching & Outreach

- 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** | Constellation Astra Fellowship

- Led a study with Owain Evans and found that some large language models (LLMs) are capable of introspection.
- Characterized the AI safety-relevant consequences of introspective models, including situational awareness.

2023 Boston **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.
- Designed and conducted a number of experiments to evaluate dataset and models.

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.

2017–2019 Berlin **Student Research Assistant** | Berlin School of Mind & Brain

Skills

Programming Python (PyTorch, scikit-learn), Javascript (node.js, jsPsych), R, C#, Unix
Dataset Creation RL Environment & Task Creation, Unity, VR, ThreeDWorld, Unix
Statistics Model Fitting & Analysis, Hypothesis Testing, Bayesian Statistics
Communication Scientific Writing, Public Science Communication, Data Visualization

Selected Publications

* indicates equal contribution.

2023 **Binder, F.**, Mattar, M., Kirsh, D., & Fan, J. Humans choose visual subgoals to reduce cognitive cost. *Proceedings of the 45th Annual Conference of the Cognitive Science Society, 7.* | 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

2021 **Binder, F.**, Mattar, M., Kirsh, D., & Fan, J. Visual scoping operations for physical assembly. *Proceedings of the 43th Annual Conference of the Cognitive Science Society, 7.* | Code & paper