Felix Jedidja Binder

E-Mail: me@felixbinder.net
URL: http://ac.felixbinder.net/

Education

2019–2025 PhD in Cognitive Science with specialization in Anthropogeny, Cognitive Science Department, University of California San Diego
Supervised by Prof. Judith Fan (Stanford), Marcelo Mattar (NYU) & David Kirsh (UCSD)

2024-2025 Visiting Researcher at Stanford University

2023 Artificial General Intelligence Safety Fundamentals Course, BlueDot Impact

2019 Diverse Intelligences Summer Institute, University of St. Andrews

2013–2019 Bachelor of Arts in Philosophy and Computer Science, Freie Universität Berlin

Experience

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*

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.
- Using Unity, created a rich 3D environment and benchmark to evaluate human and AI physical problem solving.

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 with Owain Evans | Astra Fellowship | *Constellation*

- 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.
- 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

Publications

* indicates equal contribution.

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

2024 Wang, H., Jedoui, K., Venkatesh, R., **Binder, F.**, Tenenbaum, J., Fan, J., Yamins, D., & Smith, K. Probabilistic Simulation Supports Generalizable Intuitive Physics. *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

2023 Venkatesh, R., Chen, H., Feigelis, K., Bear, D. M., Jedoui, K., Kotar, K., **Binder, F.**, Lee, W., Liu, S., Smith, K. A., Fan, J. E., & Yamins, D. L. K. Understanding Physical Dynamics with Counterfactual World Modeling. *arXiv preprint arXiv:2312.06721*.

2023 Venkatesh, R., Chen, H., Feigelis, K., Jedoui, K., Kotar, K., **Binder, F.**, Lee, W., Liu, S., Smith, K. A., Fan, J. E., & Yamins, D. L. K. Counterfactual World Modeling for Physical Dynamics Understanding. *arXiv preprint arXiv:2312.06721*.

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.*

2023 Martinez, J., **Binder, F.**, Wang, H., Haber, N., Fan, J., & Yamins, D. L. K. Measuring and Modeling Physical Intrinsic Motivation. *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

2023 Wang, H.*, Jedoui, K.*, Venkatesh, R.*, **Binder, F.***, Yamins, D., Fan, J., & Smith, K. Modeling and evaluating how the brain makes physical predictions. *Proceedings of the Society for Neuroscience.*

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: https://github.com/cogtoolslab/physics-benchmarking-neurips2021

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: https://github.com/cogtoolslab/tools_block_construction

2021 **Binder, F.***, Jones, C. R.*, Kaufman, R. A., & Lin, N. T. Cognitive cost and information gain trade off in a large-scale number guessing game. *Proceedings of the 43th Annual Conference of the Cognitive Science Society, 7.*
Code & paper: https://github.com/felixbinder/number_guessing_game

2017 **Binder, F.**, Körper als Paradigma. *cogito, München*

Prizes, Grants & Awards

2024 Career development and transition funding grant from Open Philanthropy ($15,000)

2023 Effective Ventures Long Term Future Fund Grant for study on the emergence of steganography in Large Language Models ($2,000)

2023 Won the Alignment Jam Agency Foundations Hackathon with ”Evaluating Myopia in Large Language Models” (joint work with Marco Bazzani)

2023 Cognitive Science Society Travel Award ($1,200)

2021, 2022 CARTA Fellowship in Anthropogeny ($50,000)

2019 Research grant from Templeton World Charity Foundation ($500)

Teaching

2023 COGS 18 Introduction to Programming with Dr. Eric Morgan
Teaching Assistant

2021 COGS 118B Introduction to Artificial Intelligence II with Prof. Marcelo G. Mattar
Teaching Assistant

2020-2021 COGS 100 Cyborgs Now and in the Future with Prof. David Kirsh
Teaching Assistant

2020 COGS 109 Models and Data Analysis with Prof. Megan Bardolph
Teaching Assistant

2019-2020 COGS 100 Cyborgs Now and in the Future with Prof. Taylor Scott
Teaching Assistant

Conference Presentations

2024 Panelist: COGGRAPH at *CogSci 2024* & *SIGGRAPH 2025*

2023 Talk: Best Practices for Cognitive AI Benchmarking at *Cognitive AI Benchmarking workshop at CogSci 2023*

2023 Talk: Humans Choose Visual Subgoals to Reduce Cognitive Cost at *Cognitive Science Society*

2023 Poster: Towards an Evaluation of Steganography in Large Language Models at *7th Annual Center for Human-Compatible AI (CHAI) Workshop*

2023 Poster: Humans choose visual subgoals to reduce cognitive cost at *Vision Sciences Society*

2022 Poster: Visual Scoping Operations for Physical Assembly at *Reinforcement Learning and Decision Making*

2021 Poster: *Physion: Evaluating Physical Prediction from Vision in Humans and Machines* at *NeurIPS 2021 (Datasets & Benchmarks track)*

2021 Talk: Visual scoping operations for physical assembly at *47th Annual Meeting of the Society for Philosophy and Psychology*

2021 Poster: Visual scoping operations for physical assembly at *43th Annual Conference of the Cognitive Science Society*

2021 Poster: Cognitive cost and information gain trade off in a large-scale number guessing game at *43th Annual Conference of the Cognitive Science Society*

2021 Poster: Visual Scoping for Block Construction at *CSSA* in San Diego, CA, April 2021

2018 Poster: On the boundary conditions of mind under Predictive Processing at *Open Self 2018* in Berlin, September 2018

2018 Poster: The golden ratio is not always preferred in art at *Visual Science of Art Conference 2018* Trieste, Italy, August 2018

2018 Poster: On the boundary conditions of mind under Predictive Processing at *Cognitive Systems Modelling – 7th peripatetic conference* in Małe Ciche, Poland, October 2018

2017 Talk: Körper als Paradigma at *Produktive Äquivalenz - Die Metapher im transdisziplinären Kontext*, at Humboldt Universität zu Berlin, July 2017

2017 Lecture: Körper als Paradigma zwischen Phänomenologie und Neurowissenschaften as part of lecture series *Studierendenvortrag Philosophie* at Humboldt Universität zu Berlin, May 2017

Outreach

2024-2025 Mentor at Berkeley AI Safety BASIS fellowship: mentoring 2 fellows on a project on the emergence of steganography in LLMs

2024 Talk: Can Language Models be Taught to Introspect? at *Constellation Talk Series*

2024 Talk: Language Models can be Taught to Introspect at *AI Objectives Institute*

2023 Co-founded Effective Altruism @ UC San Diego

2023 Co-organized a workshop on Cognitive AI Benchmarking at CogSci, Sydney, June 2023

2022 Outreach video for high school students on the topic ”Can Machines Think?”, San Diego, October 2022

2021 Video on perception in neural network and AI ethics for high school students with pathways2AI, San Diego, September 2021

2021 Co-organized the *2021 meeting of The Academy Neuroscience for Architecture* in San Diego, CA, September 2021

2018 Co-organized the 10th birthday of the Association of Neuroesthetics at Volksbühne Berlin, September 2018

2018 Organisation of VR body swap workshop during Watch Your Bubble Conference & exhibtion at Kunstverein Tiergarten / Galerie Nord, Berlin, May 2018

2017 Helped organize the Visual Science of Art Conference (VSAC 2017) in Berlin, August 2017

2017 Ran VR body swap workshop at Digitaler Salon at Alexander von Humboldt Institut für Internet und Gesellschaft, Berlin, November 2017

2017

Organisation of VR body swap at Digitaler Salon: Ein Herz für Cyborgs at Alexander von Humboldt Institut für Internet und Gesellschaft, Berlin, June 2017

2017 Organisation of workshop on VR body swap at Lange Nacht der Wissenschaften at Berlin School of Mind & Brain, June 2017

2017 Organisation of workshop Ich bin Du/Du bist Ich for underprivileged youth (on inducing empathy by swapping bodies in virtual reality) at Apartment Project/Schloss 19, Berlin, October – November 2017

2016 Organisation of VR body swap experiment at Digitaler Salon: Internet der Sinne at Alexander von Humboldt Institut für Internet und Gesellschaft, Berlin, November 2016

2016 Organisation of workshop Ich bin Du/Du bist Ich for underprivileged youth (on inducing empathy by swapping bodies in virtual reality) by Apartment Project/Stadtvilla Global Berlin, August 2016

Proficiency

Programming languages & frameworks

Advanced

Python
PyTorch
LaTeX

Intermediate

Java
C#
Haskell
Javascript
node.js
mongoDB
HTML & CSS
R

Scientific computing & empirical research tools

Advanced

Unix
Matlab/Octave
ThreeDWorld (TDW)
jsPsych
Pupil Labs Eyetracking System
Amazon Mechanical Turk & Prolific

Intermediate

HPC
slurm

Visual & 3D programming

Advanced

Unity
Virtual Reality
TouchDesigner
various VJ and projection mapping software
Blender
various photogrammetry software