Marcel Binz

Ingolstädter Landstraße 1, 85764 Oberschleißheim, Germany

marcel.binz@helmholtz-munich.de marcelbinz.github.io marcelbinz (+49)15203605574

Research Interests

Cognitive Science; Machine Learning; Meta-Learning; Resource Rationality; Large Language Models; Deep Learning; Bayesian Inference; Information Theory; Decision-Making; Reinforcement Learning

Position

Helmholtz Munich, Institute for Human-Centered AI **12/2023 - present**
Research scientist and deputy head

Experience

Max Planck Institute for Biological Cybernetics, PI: Dr. Eric Schulz **02/2021 - 11/2023**
Postdoctoral researcher

Harvard University, PI: Prof. Samuel Gershman **09/2019 - 12/2019**
Research visit

Facebook Inc. **06/2016 - 12/2016**
Research internship

Eberhard Karls Universität Tübingen, PI: Prof. Martin Butz **04/2015 - 08/2015**
Research assistant

Education

Philipps-Universität Marburg, PI: Prof. Dominik Endres **2018 - 2021**
Dr. rer. nat. (Psychology)

KTH Royal Institute of Technology, Stockholm **2015 - 2018**
M.Sc. (Machine Learning)

Eberhard Karls Universität Tübingen **2012 - 2015**
B.Sc. (Cognitive Science)

Peer-reviewed Publications

Note that in machine learning, unlike most other fields, conference publications are rigorously reviewed. Proceedings of selective conferences are considered archival and comparable to journals.

Demircan, C., Saanum, T., Pettini, L., **Binz, M.**, Baczkowski, B. M., Doeller, C., Garvert, M. and Schulz, E., 2024. Evaluating alignment between humans and neural network representations in image-based learning tasks. *Conference on Neural Information Processing Systems* (*NeurIPS*).

Hussain, Z., **Binz, M.**, Mata, R. and Wulff, D. U., 2024. A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*.

Coda-Forno, J., **Binz, M.**, Wang, J. X. and Schulz, E., 2024. CogBench: a large language model walks into a psychology lab. *International Conference on Machine Learning* (*ICML*).

Jagadish, A. K., Coda-Forno, J., Thalmann, M., Schulz, E. and **Binz, M.**, 2024. Human-like category

learning by injecting ecological priors from large language models into neural networks. *International Conference on Machine Learning* (*ICML*).

Schubert, J. A., Jagadish, A. K., **Binz, M.** and Schulz, E., 2024. In-context learning agents are asymmetric belief updaters. *International Conference on Machine Learning* (*ICML*).

Binz, M. and Schulz, E., 2024. Turning large language models into cognitive models. *International Conference on Learning Representations* (*ICLR*).

Binz, M., Dasgupta, I., Jagadish, A., Botvinick, M., Wang, J. X. and Schulz, E., 2023. Meta-learned models of cognition. *Behavioral and Brain Sciences* (*BBS*).

Coda-Forno, J., **Binz, M.**, Akata, Z., Botvinick, M., Wang, J. X. and Schulz, E., 2023. Meta-in-context learning in large language models. *Conference on Neural Information Processing Systems* (*NeurIPS*).

Saanum, T., Éltető, N., Dayan, P., **Binz, M.** and Schulz, E., 2023. Reinforcement learning with simple sequence priors. *Conference on Neural Information Processing Systems* (*NeurIPS*).

Schulze Buschoff, L. M., Schulz, E. and **Binz, M.**, 2023. The acquisition of physical knowledge in generative neural networks. *International Conference on Machine Learning* (*ICML*).

Binz, M. and Schulz, E., 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* (*PNAS*).

Binz, M. and Schulz, E., 2022. Reconstructing the Einstellung effect. *Computational Brain & Behavior*.

Binz, M. and Schulz, E., 2022. Modeling human exploration through resource-rational reinforcement learning. *Conference on Neural Information Processing Systems* (*NeurIPS*). **Selected as Oral**.

Binz, M., Gershman, S.J., Schulz, E. and Endres, D., 2022. Heuristics from bounded meta-learned inference. *Psychological Review*.

Brändle, F., **Binz, M.** and Schulz, E., 2022. Exploration beyond bandits. *The Drive for Knowledge: The Science of Human Information Seeking. Cambridge University Press.*

Preprints

Demircan, C., Saanum, T., Jagadish, A. K., **Binz, M.** and Schulz, E., 2024. Sparse autoencoders reveal temporal difference learning in large language models.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C., Allen, C., Schad, D., Wulff, D. U., West, J., Zhang, Q., Shiffrin, R., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z. and Schulz, E., 2023. How should the advent of large language models affect the practice of science?

Jagadish, A. K., **Binz, M.**, Saanum, T., Wang, J. X. and Schulz, E., 2023. Zero-shot compositional reinforcement learning in humans.

Coda-Forno, J., Witte, K., Jagadish, A. K., **Binz, M.**, Akata, Z. and Schulz, E., 2023. Inducing anxiety in large language models increases exploration and bias.

Non-archival Publications

Schubert, J. A., Jagadish, A. K., **Binz, M.** and Schulz, E., 2023. A rational analysis of the optimism bias using meta-reinforcement learning. *Conference on Cognitive Computational Neuroscience* (*CCN 2023*).

Jagadish, A. K., Saanum, T., Wang, J. X., **Binz, M.** and Schulz, E., 2022. Probing compositional inference in natural and artificial agents. *5th Multidisciplinary Conference on Reinforcement Learning and*

Decision Making (*RLDM 2022*).

Demircan, C., Pettini, L., Saanum, T., **Binz, M.**, Baczkowski, B. M., Doeller, C., . . . and Schulz, E., 2022. Decision-making with naturalistic options. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Binz, M. and Endres, D., 2019. Emulating human developmental stages with bayesian neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Binz, M. and Endres, D., 2019. Where do heuristics come from?. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Butz, M. V., Simonic, M., **Binz, M.**, Einig, J., Ehrenfeld, S. and Schrodt, F., 2016. Is it living? Insights from modeling event-oriented, self-motivated, acting, learning and conversing game agents. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Binz, M., Otte, S. and Zell, A., 2015. On the applicability of recurrent neural networks for pattern recognition in electroencephalography signals. *Workshop New Challenges in Neural Computation*.

Teaching

Computational cognitive science, Eberhard Karls University of Tübingen **2022, 2023**
Lecturer

International Interdisciplinary Computational Cognitive Science Summer School **2022, 2023**
Lecturer

Bayesian statistics and machine learning, Philipps-Universität Marburg **2020**
Lecturer

Theoretical neuroscience, Philipps-Universität Marburg **2019, 2020**
Lecturer

Deep learning in data science, KTH Royal Institute of Technology **2017**
Teaching assistant

Supervision (PhD students)

Julian Coda-Forno (co-supervised with Eric Schulz and Jane Wang) **2022 - present**
Meta-Learning in large language models

Akshay Kumar Jagadish (co-supervised with Eric Schulz) **2021 - present**
Reverse-engineering adaptive principles of cognition

Supervision (Master and Bachelor students)

Elif Kara **2024**
Heuristic decision-making in the wild

Johannes Schubert **2023**
Investigating the optimism bias using meta-reinforcement learning

Luca Schulze Buschoff **2022**
Development as decompression

Akshay Kumar Jagadish **2021**
Compositional generalization in meta-reinforcement learning

Gwen Hirsch **2020**
Comparing meta-learners with human performance in a continual learning framework

Hauke Niehaus **2019**
Simulating decision-making deficits in a deep meta-reinforcement-learning agent

Workshop Organization

In-context learning in natural and artificial intelligence, Rotterdam **2024**
The Annual Meeting of the Cognitive Science Society

Meta-learned models of cognition, Freiburg **2022**
The Biannual Conference of the German Cognitive Science Society

Reviewing

Nature Communications **2024 - present**
Open Mind **2024 - present**
Nature **2023 - present**
Nature Human Behaviour **2023 - present**
International Conference on Learning Representations (ICLR) **2023 - present**
Conference on Neural Information Processing Systems (NeurIPS) **2023 - present**
Behavior Research Methods **2023 - present**
Trends in Cognitive Sciences **2023 - present**
Conference on Cognitive Computational Neuroscience **2023 - present**
Proceedings of the National Academy of Sciences (PNAS) **2022 - present**
Psychological Review **2022 - present**
Computational Brain & Behavior **2022 - present**
Annual Meeting of the Cognitive Science Society **2021 - present**

Invited Talks

Research talk at Google DeepMind, London **2024**
Research talk at University of Oxford, Oxford **2024**
DGPs symposium on large language models in psychological research, Vienna **2024**
Rational altruism lab meeting, Los Angeles **2024**
Higher cognition in large language models symposium, Rotterdam **2024**
Connecting minds and machines symposium, Munich **2024**
Annual Summer Interdisciplinary Conference, Molveno **2024**
Departmental research seminar, Milan **2024**
Cognition, values & behaviour and crowd cognition joint lab meeting, Munich **2024**
Cognitive sciences colloquium, Irvine **2024**
Bosch neuro-symbolic AI focus group **2024**
Cognition, brain, & behavior research seminar, Harvard **2023**
nEuro-economics seminar series, Paris **2023**
Digital change symposium, Kloster Seeon **2023**
International Interdisciplinary Computational Cognitive Science Summer School **2023**
Language models in judgment and decision making research symposium, Vienna **2023**

Large language models meet cognitive science workshop, Sydney **2023**
Neuro-cognitive modeling group lab meeting, Tübingen **2023**
International Titisee Conference on NeuroAI, Titisee **2023**
Colloquium of the Institute of Cognitive Science, Osnabrück **2023**
Reinforcement learning and decision-making seminar, Tübingen **2023**
Conference on Neural Information Processing Systems, New Orleans **2022**
Memory, judgement and decision-making seminar, Mannheim **2022**
International Interdisciplinary Computational Cognitive Science Summer School **2022**
Conference of the German Cognitive Science Society, Freiburg **2022**
Human and machine cognition lab meeting, Tübingen **2022**
Reinforcement learning and decision-making seminar, Tübingen **2021**
Joint lab retreat: Summerfield, Schuck, Schulz **2021**
Colloquium of the Institute for Neuroinformatics, Bochum **2019**

Awards

German Cognitive Science Society Best Publication Award **2018 - 2022**
Best publication in cognitive science by a young investigator

EuroCogSci 2019 Best Poster Award **2019**
Best poster presentation

DMV-Abiturpreis **2010**
Excellent performance in high school mathematics

Grants

Google Gemma Academic Program **2024**
$5000 in credits for the Google Cloud Platform

Scientific Inference and Statistical Inference Conference **2023**
Funding for travel and accommodation

International Titisee Conference on NeuroAI **2023**
Funding for travel and accommodation

German Academic Exchange Service (DAAD) Scholarship **2019**
Funding for a three month research visit at Harvard University

Summer Institute on Bounded Rationality **2019**
Funding for travel and accommodation

Public Outreach

Science Journal for Kids version of Using cognitive psychology to understand GPT-3 **2023**
[*Link*]

Podcast on our paper Using cognitive psychology to understand GPT-3 **2023**
[*Link*]

Interview with the newspaper Tagesspiegel **2023**
[*Link*]

Talk for an event on short stories about AI hosted by the newspaper Tagblatt **2023**

[*Link*]

KFZ Science Slam **2019**

[*Link*]

Mario Lives! AAAI Video Competition winning submission **2015**

[*Link; one million views on YouTube*]