Dr. Marcel Binz
Wilhelm-Hertz-Straße 32
80805 München, Germany
marcel.binz@helmholtz-munich.de
https://marcelbinz.github.io/

Max Planck Research Group Leader: application

Dear members of the search committee,

I am writing this letter to express my interest in the "Max Planck Research Group Leader" positions offered by the MPI for Biological Cybernetics.

I am currently a research scientist and deputy head of the Institute for Human-Centered AI at Helmholtz Munich. Prior to that, I obtained a Bachelor's degree in cognitive science, a Master's degree in machine learning, a doctoral degree in psychology, and worked as a postdoctoral researcher at the Max Planck Institute for Biological Cybernetics in Tübingen.

My work is situated at the intersection of the cognitive sciences and machine learning. In particular, I use state-of-the-art machine learning methods – such as large language models and in-context learning – to uncover the fundamental principles behind human cognition. I believe that to get a full understanding of the human mind, it is vital to consider it as a whole and not just as the sum of its parts. My research goal is therefore to establish foundation models of human cognition – models that cannot only simulate, predict, and explain human behavior in a single domain but that offer a unified take on our mind. If successful that would be one of the biggest breakthroughs in the cognitive sciences.

My work has been featured in top-tier journals such as Nature, PNAS, Behavioral and Brain Sciences, and Psychological Review. In addition to that, I regularly publish at leading machine learning venues (NeurIPS, ICML, ICLR) as a first and senior author (ten main-track papers in the last two years). This interdisciplinary combination is unique and aligns perfectly with the institute's richly multidisciplinary orientation.

I have established an extensive collaboration network in both industry and academia over the recent years. From the industry side, I work closely together with my colleagues at Google DeepMind (Matthew Botvinick, Jane X. Wang, Ishita Dasgupta, Andrew K. Lampinen, Maria Eckstein, and Stephanie Chan), leading to five publications so far, several ongoing projects, and the supervision of a joint PhD student. Furthermore, I am involved in multiple collaborative efforts with academic researchers. For instance, I have recently led an international consortium with over 40 researchers (from Helmholtz Munich, the MPIs in Tübingen and Berlin, the University of Tübingen, NYU, Princeton, Google DeepMind, and others) in which we have built the first foundation model of human cognition.

I strive for excellence in mentoring and teaching, and have a proven track record in both. I am currently supervising two PhD students at Helmholtz Munich. Furthermore, I have supervised several Master's students, leading to multiple publications so far. My previous lectures have covered the topics of cognitive science, deep learning, and reinforcement learning. Most recently, I have designed a course on "Computational Cognitive Science" from scratch which I have taught twice at the Graduate Training Centre of Neuroscience in Tübingen. I am committed to continuing this trend and eager to contribute to the institute's associated educational programs.

My research was awarded the best publication award of the German Cognitive Science Society (for the best publication by a young investigator between 2018 and 2022), a DAAD Scholarships grant for a research stay at Harvard, and a selection for oral presentation at NeurIPS (rated in the top 0.2% of submitted papers). It has furthermore been featured in various media outlets, such as *New York Times*, *Spiegel*, and *Frankfurter Allgemeine Zeitung*, and been adapted into a children's version for the *Science Journal for Kids*.

To summarize, my profile ticks all the boxes required by this position, and my expertise would significantly complement and augment the research portfolio at the MPI for Biological Cybernetics. I would be excited to bring back my competences, skills, and network to an outstanding institution like yours and shape its future development. Thank you for taking the time to consider my application. I am looking forward to hearing from you.

Kind regards,
Marcel Binz

Humans are remarkable generalists. Not only do we routinely make mundane decisions, like choosing a breakfast cereal or selecting an outfit, but we also tackle complex challenges, such as figuring out how to cure cancer or how to explore outer space. We learn new skills from just a few demonstrations, we engage in causal reasoning, and we are curious. We climb mountains, we spend hours playing video games, and we create captivating art.

My research employs state-of-the-art machine learning methods to uncover the fundamental principles behind these abilities. I believe that to get a full understanding of the human mind, it is vital to consider it as a whole and not just as the sum of its parts. My research goal is therefore to establish **foundation models of human cognition** – models that cannot only simulate, predict, and explain human behavior in a single domain but that offer a unified take on our mind.

The importance of such a unified approach has already been recognized by the pioneers of cognitive science research. For example, in 1990, Newell stated "unified theories of cognition are the only way to bring [our] wonderful, increasing fund of knowledge under intellectual control" [Newell'90]. Yet, several decades later, we are still far away from such a theory. Why is that the case? It turned out that hand-engineering a single computational model of the entire human mind is a challenging – perhaps even impossible – endeavor. My work takes a radically different approach towards establishing a unified theory of cognition. It is based on **data-driven principles** and utilizes **state-of-the-art machine learning methods** – such as large language models and in-context learning – to understand the human mind at scale. If successful that would be one of the biggest breakthroughs in the cognitive sciences.

Summaries of scientific achievements

Large language models and cognitive science

Large language models (LLMs) are taking our society by storm. While these models show many exceptional abilities, they are known to be notoriously uninterpretable. I believe that one approach toward understanding LLMs may come from cognitive science – after all cognitive scientists have already been trying to understand another immensely complex system for decades: the human mind. My research was amongst the first to demonstrate that modern LLMs can be assessed using tools from psychology. For example, I have tested a particular LLM (GPT-3) on a battery of canonical experiments from the psychology literature, including tasks that require decision-making, information search, deliberation, and causal reasoning [PNAS'23]. While GPT-3's overall performance in solving these tasks was impressive, I found that the model showed no signatures of directed exploration and that it failed miserably in a causal reasoning task, both of which are problems that require active engagement with one's environment. These results suggest direct ways to improve the next generation of LLMs. This work was published in PNAS, adapted into a kids' version for the Science Journal for Kids, and featured in various media outlets. Together with a group of researchers from Google DeepMind, I have recently written a review article that lays the conceptual groundwork for this new and emerging field of *machine psychology* [arXiv'24a]. Furthermore, I am supervising several students who are continuously extending this line of research [ICML'24a; ICML'24b; ICML'24c; NeurIPS'23a; NeurIPS'24; ICLR'25; arXiv'23a].

While cognitive science is valuable for understanding and improving LLMs, the interaction also goes the other way around – with LLMs being used as tools to inform our theories of human cognition [ICLR'24]. I have recently initiated a large-scale collaboration with over 40 researchers from across the globe to build the first foundation model of human cognition. This project resulted in a model called Centaur that predicts human behavior better than existing domain-specific cognitive models in almost every single setting [Nature'25]. Furthermore, Centaur is able to generalize to new cover stories, problem structures, and even entirely novel domains. While mere access to such a model will already revolutionize the field of psychology, I want to now use it to reverse-engineer human cognition at an unprecedented scale as discussed in more detail in my future research plans. In the process of building this model, I have also created the largest cross-domain data set of human behavior in existence – Psych-101 – which covers over 10,000,000 choices from over 60,000 participants in 160 different experiments. Psych-101 enables totally new types of research and will be a valuable resource for both the cognitive science and the machine learning community.

Finally, it is important to me to consider the broader impact that LLMs have on science. I have recently led an interdisciplinary effort on the question of "how should the advancement of LLMs affect the practice of science?" [PNAS'25]. An issue that is close to my heart in this context is that we need to move away from proprietary models and increase the reliance on open-source models. To democratize the use of such models, I have co-authored a tutorial outlining how to use open-source LLMs in the behavioral sciences [BRM'24].

Meta-learned models of cognition

In-context learning – the ability to learn from examples presented in a model's context – is at the heart of LLMs. Yet, LLMs are not the only models that have this ability. It can also emerge in sequence models that are trained through repeated encounters with an environment. This type of learning is commonly referred to as meta-learning. My work established a connection between in-context learning systems that arise through meta-learning and human cognition. In particular, I have shown that in-context learning can explain behavioral phenomena across a wide range of domains, including decision-making [PsychReview'22], exploration [NeurIPS'22], category learning [ICML'24c], compositional reasoning [arXiv'23b], curriculum learning [PhD'21], and others [ICML'24b].

Together with my colleagues at Google DeepMind, I have recently summarized my vision of this emerging research program in an article published at Behavioral and Brain Sciences [BBS'23]. In this article, I point out that in-context learning combines the advantages of two of the most successful modeling frameworks in cognitive science: it is as scalable and flexible as neural network models but – like Bayesian models – it can also be equipped with prior knowledge that facilitates sample efficient learning. This powerful combination makes it possible to lift cognitive modeling to more naturalistic settings that are out of reach for traditional frameworks.

Research plans

There will be a paradigm shift in cognitive science towards more unified models in the near future. I want to use recent advances from machine learning to build such models. This will allow me to reverse-engineer human cognition at an unprecedented scale. In the following, I outline three research directions that exemplify this mission.

My main goal is to bring cognition back into Centaur using a theory-driven approach, thereby transforming the behaviorally predictive model into a unified theory of cognition. This requires systematically integrating cognitive processes into the model's architecture, transitioning it to Marr's algorithmic level and allowing me to address pivotal questions such as "which neural architecture offers the best description of human behavior?" In this context, I plan to focus on three key aspects of human cognition: (1) memory, (2) learning, and (3) planning, mapping onto the questions "how is memory structured?", "how is new information integrated?" and "how is knowledge transformed into behavior?" Even though my research will focus on using the resulting models to improve our understanding of human cognition, it also offers the potential to identify new neural network architectures that could lead to better and more robust machine learning systems.

I intend to supplement this theory-driven approach with discovery-driven approaches that can uncover human cognitive processes *without* a priori postulating their existence. Modern state-of-the-art algorithms from neuroscience and mechanistic interpretability enable me to conduct this kind of research. I plan to use such methods to poke at the internal mechanisms of black-box foundation models of models like Centaur and extract interpretable structures that can be used to generate novel hypotheses about human information processing. This approach is particularly powerful in domains where we do not have a solid grasp on human cognition yet, and allows me to target questions such as "what intermediate steps do people perform when solving a particular problem?" or "how are cognitive processes in different domains connected with each other?" There are already a wide range of tools just waiting for someone to explore them for this purpose. The direction that I am perhaps most excited about is to identify interpretable neurons and computational circuits using sparse autoencoders. In particular, I want to use this approach to find a semantically meaningful interpretation for every single neuron in Centaur, thereby building a cognitive atlas of it. For a given record of an experimental session, it would then be possible to identify which neurons are activated, and use this atlas to make inferences about the underlying cognitive processes of the individual who generated the data.

True foundation models of human cognition do not only need to solve text-based problems but must also process high-dimensional visual information and navigate naturalistic environments. Building such systems requires access to the right data. While we have a lot of text data to train language models, we currently lack large-scale, open-source data sets of human behavior in naturalistic environments. At present, such data is almost exclusively collected in industry labs. I want to change that and collect and gather the largest multi-modal data set of people using different tools and interacting with each other. This data set will enable me to conduct unprecedented studies on human cognition in naturalistic environments.

Project- and subject-related list of publications

[arXiv'23a] Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, **Marcel Binz**, Zeynep Akata, and Eric Schulz. "Inducing anxiety in large language models increases exploration and bias". In: *arXiv preprint*. 2023.

[arXiv'23b] Akshay Kumar Jagadish, **Marcel Binz**, Tankred Saanum, Jan X Wang, and Eric Schulz. "Zero-shot compositional reinforcement learning in humans". In: *arXiv preprint*. 2023.

[arXiv'24a] Thilo Hagendorff, Ishita Dasgupta, **Marcel Binz**, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. "Machine Psychology". In: *arXiv preprint*. 2024.

[BBS'23] **Marcel Binz**, Ishita Dasgupta, Akshay K Jagadish, Matthew Botvinick, Jane X Wang, and Eric Schulz. "Meta-learned models of cognition". In: *Behavioral and Brain Sciences* (2023), pp. 1–38.

[BRM'24] Zak Hussain, **Marcel Binz**, Rui Mata, and Dirk U Wulff. "A tutorial on open-source large language models for behavioral science". In: *Behavior Research Methods* (2024), pp. 1–24.

[CoBB'22] **Marcel Binz** and Eric Schulz. "Reconstructing the Einstellung effect". In: *Computational Brain & Behavior* (2022), pp. 1–17.

[ICLR'24] **Marcel Binz** and Eric Schulz. "Turning large language models into cognitive models". In: *International Conference on Learning Representations (ICLR)*. 2024.

[ICLR'25] Can Demircan, Tankred Saanum, Akshay K. Jagadish, **Marcel Binz**, and Eric Schulz. "Sparse Autoencoders Reveal Temporal Difference Learning in Large Language Models". In: *International Conference on Learning Representations (ICLR)*. 2025.

[ICML'23] Luca Schulze Buschoff, Eric Schulz, and **Marcel Binz**. "The Acquisition of Physical Knowledge in Generative Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2023.

[ICML'24a] Julian Coda-Forno, **Marcel Binz**, Jane X Wang, and Eric Schulz. "CogBench: a large language model walks into a psychology lab". In: *International Conference on Machine Learning (ICML)*. 2024.

[ICML'24b] Johannes A. Schubert, Akshay Kumar Jagadish, **Marcel Binz**, and Eric Schulz. "In-Context Learning Agents Are Asymmetric Belief Updaters". In: *International Conference on Machine Learning (ICML)*. 2024.

[ICML'24c] Akshay Kumar Jagadish, Julian Coda-Forno, Mirko Thalmann, Eric Schulz, and **Marcel Binz**. "Human-like Category Learning by Injecting Ecological Priors from Large Language Models into Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2024.

[Nature'25] **Marcel Binz** et al. "A foundation model to predict and capture human cognition". In: *Nature*. 2025.

[NeurIPS'22] **Marcel Binz** and Eric Schulz. "Modeling Human Exploration Through Resource-Rational Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[NeurIPS'23a] Julian Coda-Forno, **Marcel Binz**, Zeynep Akata, Matthew Botvinick, Jane X Wang, and Eric Schulz. "Meta-in-context learning in large language models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[NeurIPS'23b] Tankred Saanum, Noémi Éltető, Peter Dayan, **Marcel Binz**, and Eric Schulz. "Reinforcement Learning with Simple Sequence Priors". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[NeurIPS'24] Can Demircan, Tankred Saanum, Leonardo Pettini, **Marcel Binz**, Blazej M Baczkowski, Paula Kaanders, Christian F Doeller, Mona M Garvert, and Eric Schulz. "Evaluating alignment between humans and neural network representations in image-based learning tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[Newell'90] Allen Newell. *Unified Theories of Cognition*. Cambridge: Harvard University Press, 1990.

[PhD'21] **Marcel Binz**. "Principles of Human Learning". PhD thesis. Philipps-Universität Marburg, 2021.

[PNAS'23] **Marcel Binz** and Eric Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120.

[PNAS'25] **Marcel Binz** et al. "How should the advancement of large language models affect the practice of science?" In: *Proceedings of the National Academy of Sciences* (2025).

[PsychReview'22] **Marcel Binz**, Samuel J Gershman, Eric Schulz, and Dominik Endres. "Heuristics from bounded meta-learned inference." In: *Psychological review* 129.5 (2022), p. 1042.