ERC Starting Grant 2026

Part B1

Modeling the Integrated minD At Scale

MIDAS

Cover Page:

Name of the Principal Investigator (PI)	Marcel Binz
Name of the PI's Host Institution for the project	Helmholtz Zentrum München German Research Center for Environmental Health
Proposal duration in months	60

At the heart of cognitive science lies a fundamental question: how does the human mind transform incoming information into behavior? Computational cognitive models translate psychological theories into algorithmic form, making it possible to uncover cognitive processes that are not directly observable in behavior alone. Yet, despite their vast potential for advancing our understanding of the human mind, current cognitive models remain largely limited to narrow tasks or domains. This fragmentation has contributed to psychology's broader generalization crisis and hindered theoretical integration. Modeling the Integrated minD At Scale (MIDAS) tackles this challenge by establishing a fundamentally new modeling paradigm – one centered around large-scale, integrated models capable of simulating, predicting, and explaining human behavior across a wide range of experimental paradigms. In doing so, MIDAS marks a transformative step for cognitive modeling, shifting the field from fragmented, task-specific approaches toward a unified theory that reflects the full richness of human cognition. MIDAS is structured around three key objectives. First, to develop cognitive models capable of capturing human behavior in any experiment expressible in natural language. Second, to establish a robust methodological framework for evaluating and comparing large-scale, integrated models. Third, to extend such models to more naturalistic environments. To achieve these objectives, MIDAS draws on cutting-edge methods across cognitive science, machine learning, and language modeling, applied to behavioral data at unprecedented scale. My prior work has laid the foundations for this approach, notably through my leadership of an international collaboration that produced the largest cross-domain dataset of human behavior to date. With the methodological tools and infrastructure now sufficiently mature, the field is ready for a transformation – and I am uniquely positioned to spearhead it.

Part I of the Scientific Proposal (max. 5 pages, references do not count toward the page limit).

At the heart of cognitive science lies a fundamental question: how does the human mind transform incoming information into behavior? **Computational cognitive models** translate psychological theories into algorithmic form, making it possible to uncover cognitive processes that are not directly observable in behavior alone [1, 2, 3]. Yet, despite their vast potential for advancing our understanding of the human mind, the current landscape of cognitive modeling remains narrowly scoped and domain-specific: almost every experimental paradigm gives rise to its own idiosyncratic model. The resulting fragmentation of theories has contributed to psychology's broader generalization crisis and continues to hinder theoretical integration [4]. Importantly, this issue is more than a minor inconvenience but rather a central bottleneck holding back progress in the field.

What the field desperately needs is a single, unified model that can engage with any experiment – be it decision-making [5, 6], reinforcement learning [7, 8, 9], or problem-solving [10] – and produce not only human-like responses but also interpretable, theory-grounded explanations of human behavior. **M**odeling the **I**ntegrated min**D** **A**t **S**cale (**MIDAS**) realizes this vision by constructing an integrated account of human cognition. To accomplish this, MIDAS integrates state-of-the-art methods from cognitive science, machine learning, and language modeling and applies them to behavioral data at unprecedented scale. In doing so, MIDAS sets out to shift the study of the human mind from a collection of fragmented, task-specific approaches toward a unified theory that captures the full richness of human cognition.

Why now?

As early as 1990, Alan Newell argued that “unified theories of cognition are the only way to bring [our] wonderful, increasing fund of knowledge under intellectual control” [11]. Yet, decades later, we are still far away from such a theory, raising the question: why has this goal remained so elusive? One major reason is that hand-designing a single model that captures the entire complexity of the human mind has proven to be a challenging – perhaps even impossible – endeavor. **MIDAS takes a radically different approach towards establishing a unified theory of cognition [12, 13]. It is based on data-driven principles and utilizes state-of-the-art machine learning methods to understand the human mind at scale [14, 15].**

The key ingredients – **large-scale behavioral datasets**, **expressive computational models**, **and high-performance compute infrastructures** – are now finally in place for realizing this vision, thanks to the groundwork laid in my prior work:

1. For the first time, large-scale behavioral datasets make it possible to model human cognition across tasks and domains. More specifically, MIDAS is powered by **Psych-101** – a behavioral dataset comprised of over 10,000,000 individual choices across 160 experiments, all standardized into a unified format that enables integrative modeling [16, 17]. This effort was the result of a large-scale international collaboration that I initiated and led, involving over 40 researchers from institutions including Princeton, NYU, Google DeepMind, and multiple Max Planck Institutes – demonstrating both the feasibility and the broad community support for this vision.
2. My prior work has demonstrated that this dataset can be used to fine-tune a large language model, leading to a computational model called **Centaur** that closely *mimics* human behavior across a wide range of domains [16]. Centaur is a powerful predictive system: it can anticipate human choices with high accuracy across a wide range of tasks. However, in its current form, it offers limited insight into the underlying cognitive processes. It remains a black-box model – useful for many applications, but not explanatory in itself [18]. MIDAS will take the crucial next step: developing models that are not only predictive, but also *explanatory*.
3. Finally, recent advances in compute infrastructure have drastically reduced the time required to train large-scale computational models. For instance, training a GPT-2-scale model [19] can now be accomplished in under five minutes on a small-scale local GPU cluster [20]. This enables rapid iteration over model architectures and hypotheses – a key requirement for cognitive modeling at the scale of MIDAS.

MIDAS leverages these pillars to develop computational models that not only predict behavior, but also reveal the computational principles of human cognition at an unprecedented scale.

Objectives

MIDAS is structured around three key objectives:

1. To develop integrated, interpretable and theory-grounded models of the human mind.
2. To establish a methodological framework for evaluating large-scale cognitive models.
3. To move cognitive modeling to more naturalistic environments.

Figure 1 summarizes these objectives and their relation to the three corresponding work packages (WPs) described below.

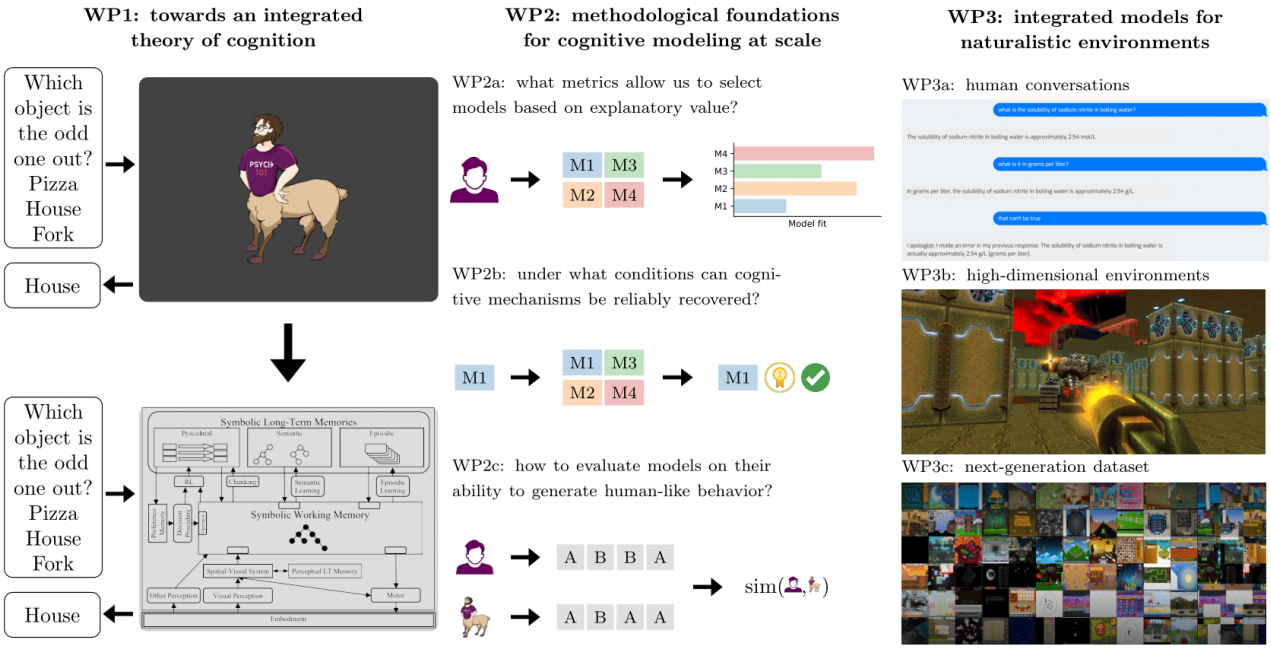

Figure 1. Illustration of the different work packages of MIDAS.

WP1: towards an integrated theory of cognition

Objective: To develop integrated, interpretable and theory-grounded models of the human mind.

The goal of WP1 is to develop large-scale, integrated models of the human mind by combining the expressive power of large language models with the theoretical grounding of cognitive science. The resulting hybrid models will be both highly predictive and structurally interpretable [21] – able to simulate human behavior in any experimental paradigm that can be described in natural language, while also shedding light on the underlying computational mechanisms of cognition. WP1 will then use these models to address key questions such as:

- Which model architecture best captures human behavior across a wide range of domains?
- Which cognitive processes are domain-general, and which are domain-specific?
- How do distinct cognitive processes interact with each other to produce behavior?

It is now clear that large language models can achieve high predictive accuracy across a wide range of cognitive tasks [16]. Yet despite their predictive success, such models are typically black-box systems, offering limited insight into the cognitive processes they emulate. WP1 addresses this limitation by embedding structured assumptions about **memory**, **learning**, and **planning** into those models, thereby transforming them from predictive tools into explanatory theories of cognition [22].

More specifically, WP1 will develop models that implement distinct assumptions about core cognitive processes. These models will build on the transformer architecture as a backbone but extend it with human-like memory systems [23, 24, 25], learning rules [26, 27], and planning capabilities [28]. Transformers offer a natural starting point for this approach: they are powerful enough to fit human

behavior, as shown in my work on Centaur [16], and their structure resonates with theories in neuroscience and cognitive science [29, 30, 31, 32]. WP1 will build on this foundation and go beyond it by harnessing advances from modern deep learning, incorporating mechanisms from cognitive modeling, and developing new methods where needed. Each candidate model will be fitted to behavioral data from Psych-101 [16] and evaluated on its ability to capture human behavior. **WP1 will thus pioneer the first generation of large-scale, integrated models of cognition, marking a paradigm shift in the scientific study of the human mind.**

WP2: methodological foundations for cognitive modeling at scale

Objective: To establish a methodological framework for evaluating large-scale cognitive models.

The models developed in WP1 will be evaluated using well-established tools from the cognitive modeling literature, including parameter fitting, model comparisons, and posterior predictive checks [1]. However, unlike conventional cognitive models, which typically involve only a handful of parameters, the models developed in WP1 are expected to be orders of magnitude larger. While the aforementioned methods are the gold-standard in conventional cognitive modeling, they have never been applied at the scale and complexity envisioned by MIDAS. WP2 closes this gap by building the computational, statistical, and theoretical foundations required for the evaluation of large-scale, integrated models of cognition.

This work package will focus on three key questions:

1. **What metrics allow us to evaluate large-scale, integrated models not just for prediction, but for explanatory value?** When neural network models are fitted to human data, their performance is typically evaluated using cross-validation on held-out data [14, 15, 16]. While effective for assessing predictive accuracy, this approach tends to favor highly flexible models at the expense of interpretability and explanatory power. Bayesian model comparison offers a principled alternative, with formal guarantees for identifying the true data-generating process under ideal conditions [33, 34]. Yet its application to neural network models remains virtually unexplored in the context of human cognition, primarily due to the intractability of exact inference in high-dimensional settings. WP2 explores a range of approximation methods for Bayesian model comparison in neural networks [35], aiming to make it the default tool for large-scale, integrative modeling.
2. **Under what conditions can cognitive mechanisms in large-scale, integrated models be reliably inferred from behavioral data?** We want to ensure that MIDAS enables trustworthy inferences about the cognitive processes underlying human behavior. This requires knowing when the mechanisms inside a model can be meaningfully recovered from behavioral data [1]. To identify these conditions, WP2 will conduct extensive model simulation and recovery studies: generating behavior from models with known mechanisms and testing whether those mechanisms can be accurately recovered. This approach will make it possible to delineate the conditions under which reliable inferences are possible and when additional constraints are needed.
3. **How can we evaluate models not only in terms of predictive performance, but also their ability to generate human-like behavior?** A good predictive fit to human data does not guarantee that a model generates human-like behavior [36, 37]. This disconnect becomes increasingly pronounced as models grow more expressive. To address this issue, additional forms of model evaluation – such as posterior predictive checks – are needed [1]. Posterior predictive checks ask, in essence, whether a model, when used to generate behavior, can reproduce the qualitative patterns observed in empirical data. WP2 aims to develop such posterior predictive checks for large-scale, integrated models of cognition. Its contributions are two-fold: (a) the development of a benchmark for simulating and evaluating models against a broad set of established cognitive effects, and (b) the design of a theory-agnostic method for conducting posterior predictive checks at scale. The outcome will be a set of principled methods for determining whether large-scale, integrated models genuinely capture the structure of human behavior.

Together, these advances will establish the theoretical and methodological foundations required for cognitive modeling at unprecedented scale. WP2 will run in parallel to WP1 to

achieve maximal synergistic effects and serve as a link between model development and application to naturalistic environments (WP3).

WP3: integrated models for naturalistic environments

Objective: To move cognitive modeling to more naturalistic environments.

While most cognitive models are designed for controlled laboratory settings, real-world cognition unfolds in rich, multimodal environments [38, 39]. A truly integrated model of human cognition must thus ultimately account for behavior in such environments. WP3 addresses this challenge by extending the insights gained in WP1 and WP2 to more naturalistic environments, thereby supporting model evaluation under conditions that require integrated decision-making, interaction, and perception.

WP3 consists of three main components, each targeting a key dimension of naturalistic cognition: more complex behavioral sequences, more complex visual inputs, and the development of a novel, large-scale dataset of people engaging in naturalistic, computer-based tasks. Each component is outlined in turn below.

1. **Modeling human conversations.** Human conversation is one of the most information-rich forms of behavior. It requires the integration of memory, perception, reasoning, and social inference, all within real-time, context-sensitive exchanges [40, 41]. Furthermore, unlike traditional laboratory tasks – which typically involve choosing from a discrete set of predefined responses – conversation is truly open-ended. WP3 aims to develop the first full-scale cognitive model of human dialogue, made possible by the advances in WP1. The goals are twofold: (1) to determine how much we can infer about core cognitive processes purely from conversational data, and (2) to uncover how language interfaces with other cognitive processes such as memory, reasoning, and decision-making. The results could show that language offers a uniquely rich signal for inferring cognitive processes – or clarify its limits in doing so [42]. Either outcome would be foundational.
2. **Large-scale cognitive models for high-dimensional sensorimotor environments.** While much progress has been made in modeling human perception [43, 44], few cognitive models have addressed how perception and action are coupled in high-dimensional sensorimotor environments [45, 46]. WP3 addresses this gap by developing large-scale cognitive models for a real-time 3D game environment that require motor control, spatial reasoning, and long-term planning. The resulting models will be grounded in the architectures developed in WP1 and extended to handle rich perceptual input, thereby providing a unique opportunity to investigate how action shapes perception and vice versa. In doing so, WP3 will enable the empirical evaluation of competing hypotheses about how the brain processes visual input: whether perception is largely feedforward and task-agnostic, jointly optimized with behavior, or shaped by higher-level cognitive goals. A fully integrated model for such settings would mark a turning point in cognitive modeling and establish a new standard for future research.
3. **Towards a novel, large-scale dataset for naturalistic, computer-based tasks.** Finally, WP3 will lay the groundwork for a new kind of dataset: one that captures human behavior across a diverse range of naturalistic, computer-based tasks, including games, interactive tool use, and behavioral experiments. To initiate this effort, WP3 will convene a dedicated workshop bringing together leading researchers in cognitive science, machine learning, and human-computer interaction to develop a technical and organizational blueprint for this effort. While WP3 will produce a concrete prototype and coordination plan, its primary goal is to establish the foundation for a future multi-lab initiative aimed at building the next-generation dataset for modeling human cognition in naturalistic settings.

Together, these components push cognitive modeling beyond controlled laboratory settings, with the long-term goal of developing a model capable of performing any computer-based task in a human-like manner, while offering deep insights into the architecture of human cognition. **MIDAS thus marks the beginning of a new era in cognitive modeling – one defined by models operating in rich, naturalistic environments.**

Impact

MIDAS redefines the landscape of cognitive science by establishing a large-scale, integrative, and interpretable modeling framework for human cognition. Its impact spans multiple dimensions: scientific, by advancing unified theories of the mind; methodological, by setting new standards for model evaluation at scale; and applied, by enabling a transition toward naturalistic environments that pave the way for future real-world use cases.

Scientific impact. Psychology is currently grappling with a generalization crisis, in which many findings fail to extend beyond the narrowly defined experimental settings in which they were originally discovered [4]. Overcoming this crisis requires integrating knowledge across tasks, domains, and modalities. MIDAS rises to this challenge by developing large-scale, integrated models of cognition grounded in advances from cognitive science [3, 6], machine learning [14, 15], and language modeling [16, 19, 47]. This effort will culminate in a set of computational models capable of simulating human behavior across a wide range of experimental tasks, while also providing interpretable, theory-driven explanations of the underlying cognitive processes (see WP1). This type of integrative modeling aims to reveal answers to pivotal questions of cognitive science that, until now, have remained out of reach, such as which cognitive mechanisms are shared across domains, which are task-specific, and how they interact within a unified system.

Methodological impact. MIDAS will set new standards for how large-scale, integrated models of cognition are built, evaluated, and compared. Whereas cognitive models have traditionally been validated in narrow, small-scale experimental settings, MIDAS both pushes the limits of existing methods and develops new ones where needed (see WP2). In doing so, it establishes a robust methodological foundation for cumulative, transparent, and reproducible science. The resulting tools and datasets will be openly shared, creating a lasting resource for the cognitive science community.

Applied impact. Bringing cognitive models closer to the complexity of real-world behavior is desperately needed to unlock their full practical potential. MIDAS's expansion into naturalistic environments (see WP3) marks a critical step in this direction and thereby paves the way for future applications in education, psychiatry, and political discourse. As such, the project stands to deliver one of the most significant breakthroughs in cognitive science, establishing a new generation of models that are both theoretically grounded and practically impactful.

Conclusion

Thirty-five years after Newell's call for a unified theory of human cognition, this vision is now within reach [11]. For the first time, we have all the necessary ingredients in place: rich behavioral data, powerful computational models, and the resources to discover such models in a scalable, data-driven way. With my interdisciplinary expertise at the intersection of cognitive science, machine learning, and language modeling – and a proven track record of leading large-scale, collaborative initiatives – I am uniquely positioned to bring this vision to life. ERC funding will not only make this project possible but catalyze a long-overdue transformation in cognitive science: from fragmented, task-specific models toward a unified theory of the mind.

What is MIDAS? MIDAS aims to build large-scale, integrated models of human cognition that can simulate, predict, and explain human behavior across a broad range of tasks and environments.

Why is it impactful? By moving beyond fragmented, domain-specific models, MIDAS tackles the generalization crisis in psychology and paves the way for a unified theory of human cognition.

How? MIDAS integrates state-of-the-art methods from cognitive science, machine learning, and language modeling, applying them to large-scale behavioral data to yield theory-grounded explanations of cognition.

Why me? My unparalleled expertise in cognitive modeling and machine learning, together with proven experience in leading large-scale projects, uniquely equips me to execute this vision.

References

[1] Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *elife*, *8*, e49547.

[2] McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11-38.

[3] **Binz, M.**, Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147.

[4] Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.

[5] **Binz, M.**, Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological review*, *129*(5), 1042.

[6] **Binz, M.**, & Schulz, E. (2024). Turning large language models into cognitive models. In *Twelfth International Conference on Learning Representations (ICLR)*.

[7] **Binz, M.**, & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. *Advances in neural information processing systems*, *35*, 31755-31768.

[8] Demircan, C., Saanum, T., Pettini, L., **Binz, M.**, Baczkowski, B., Doeller, C., ... & Schulz, E. (2024). Evaluating alignment between humans and neural network representations in image-based learning tasks. *Advances in Neural Information Processing Systems*, *37*, 122406-122433.

[9] Schubert, J. A., Jagadish, A. K., **Binz, M.**, & Schulz, E. (2024). In-context learning agents are asymmetric belief updaters. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 43928-43946).

[10] **Binz, M.**, & Schulz, E. (2023). Reconstructing the Einstellung effect. *Computational Brain & Behavior*, *6*(3), 526-542.

[11] Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

[12] Rosas, F. E., Luppi, A. I., Mediano, P. A., Kringelbach, M. L., Pessoa, L., & Turkheimer, F. (2025). Top-down and bottom-up neuroscience: overcoming the clash of research cultures. *Nature Reviews Neuroscience*, 1-3.

[13] Baggio, G. (2025). Could machine learning help to build a unified theory of cognition?.

[14] Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209-1214.

[15] Ji-An, L., Benna, M. K., & Mattar, M. G. (2025). Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 1-9.

[16] **Binz, M.**, Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... & Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, 1-8.

[17] **Binz, M.**, & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

[18] Bowers, J., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. (2025). Centaur: A model without a theory.

[19] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[20] Keller, J., Bernstein, J., Rappazzo, B., fernbear.bsky.social, Vlado, B., Jiacheng, Y., Cesista, F., Koszarsky, B., & Grad62304977. (2024). modded-nanogpt: Speedrunning the NanoGPT baseline. GitHub. https://github.com/KellerJordan/modded-nanogpt

[21] Eckstein, M., Summerfield, C., Daw, N., & Miller, K. J. (2024). Hybrid Neural-Cognitive Models Reveal How Memory Shapes Human Reward Learning.

[22] Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

[23] Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

[24] Tulving, E. (1972). Episodic and semantic memory. *Organization of memory*, *1*(381-403), 1.

[25] Tulving, E. (1985). How many memory systems are there?. *American psychologist*, *40*(4), 385.

[26] Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, *2*, 64-69.

[27] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

[28] Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

[29] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., ... & Hochreiter, S. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

[30] Whittington, J. C., Warren, J., & Behrens, T. E. (2021). Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*.

[31] Li, J. A., Zhou, C., Benna, M., & Mattar, M. G. (2024). Linking in-context learning in transformers to human episodic memory. *Advances in neural information processing systems*, *37*, 6180-6212.

[32] Gershman, S. J., Fiete, I., & Irie, K. (2025). Key-value memory in the brain. *Neuron*, *113*(11), 1694-1707.

[33] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

[34] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[35] Immer, A., Bauer, M., Fortuin, V., Rätsch, G., & Emtiyaz, K. M. (2021, July). Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning* (pp. 4563-4573). PMLR.

[36] Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, *21*(6), 425-433.

[37] Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current opinion in behavioral sciences*, *11*, 49-54.

[38] Wise, T., Emery, K., & Radulescu, A. (2024). Naturalistic reinforcement learning. *Trends in Cognitive Sciences*, *28*(2), 144-158.

[39] Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.

[40] Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

[41] Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

[42] Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*(8017), 575-586.

[43] Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413-423.

[44] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.

[45] Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*(3), 425-442.

[46] Raad, Maria Abi, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield et al. "Scaling instructable agents across many simulated worlds." *arXiv preprint arXiv:2404.10179* (2024).

[47] Hussain, Z., **Binz, M.**, Mata, R., & Wulff, D. U. (2024). A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, *56*(8), 8214-8237.

Curriculum vitae and Track Record (max. 4 pages)

PERSONAL DETAILS

Family name, First name:	Binz, Marcel
ORCID:	https://orcid.org/0000-0001-8872-8386
URL for website:	https://marcelbinz.github.io/
GitHub:	https://github.com/marcelbinz/

EDUCATION

20/04/2021 **Dr. rer. nat. in Psychology**
Department of Psychology, Philipps-Universität Marburg, Germany
Advisor: Prof. Dominik Endres
Thesis: Principles of Human Learning
Developed a new framework linking meta-learning and in-context learning to human cognition.

2018 **M.Sc. in Machine Learning**
Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Sweden
Advisor: Prof. Florian Pokorny
Thesis: Learning Goal-Directed Behaviour

2015 **B.Sc. in Cognitive Science**
Department of Computer Science, Eberhard Karls Universität Tübingen, Germany
Advisor: Prof. Sebastian Otte
Thesis: Pattern Recognition in Electroencephalography Signals with Recurrent Neural Networks

CURRENT POSITIONS

2023 – **Research scientist and deputy head**
Institute for Human-Centered AI, Helmholtz Munich, Germany
in the institute of Dr. Eric Schulz
Initiated and led a large-scale international collaboration to create the first foundation model of human cognition.

PREVIOUS POSITIONS

2021 – 2023 **Postdoctoral scientist**
Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, Germany
in the group of Dr. Eric Schulz
Pioneered groundbreaking research applying cognitive science principles to understand large language models.

2019 **Visiting researcher**
Department of Psychology, Harvard University, United States
in the lab of Prof. Samuel Gershman

2016 **Research intern**
Facebook Inc., San Francisco, United States

RESEARCH ACHIEVEMENTS AND PEER RECOGNITION

I have published over twenty papers across cognitive science and machine learning in the

recent years. My work has appeared in top-tier journals such as Nature, PNAS, Behavioral and Brain Sciences, and Psychological Review. In addition, I regularly publish in leading machine learning conferences, including NeurIPS, ICML, and ICLR. This rare interdisciplinary profile uniquely positions me to lead the present project. A subset of these publications is listed below.

1. **Binz, M.**[*], Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... & Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, 1-8.
   Project initiator and lead coordinator. Built the largest behavioral dataset and the first foundation model for cognition. The data and methods established in this project set the stage for the present proposal.

2. **Binz, M.**[*], & Schulz, E. (2024). Turning large language models into cognitive models. In *Twelfth International Conference on Learning Representations (ICLR)*.
   Demonstrated how large language models can be finetuned to mimic human behavior. This project laid the groundwork for the subsequent Nature paper above.

3. **Binz, M.**[*], & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.
   A seminal study launching a new subfield in the evaluation of large language models through cognitive psychology. This project started a whole new line of research, eventually cumulating in the present proposal.

4. **Binz, M.**[†,*], Alaniz, S.[†], Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., ... & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, *122*(5), e2401227121.
   Led and coordinated an adversarial collaboration exploring how large language models should be used in scientific practice. The resulting manuscript discusses key ethical and societal implications, many of which are directly relevant to the present proposal.

5. **Binz, M.**[*], Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147.
   This publication has established itself as a central reference for applying meta-learning to cognitive modeling. Its influence is underscored by over twenty invited commentaries from leading researchers.

6. **Binz, M.**, Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological review*, *129*(5), 1042.
   We show that different heuristic decision-making strategies emerge from (a) adaptation to the environment and (b) constraints on computational resources. The resulting models make precise predictions on which decision-making strategies people use that were verified in three behavioral studies.

7. **Binz, M.**[*], & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. *Advances in neural information processing systems*, *35*, 31755-31768.
   We show that human exploration is well explained by the principle of resource rationality. The resulting models outperform previous approaches in capturing human exploratory behavior and align with findings from developmental and clinical studies.

8. Jagadish, A. K., Coda-Forno, J., Thalmann, M., Schulz, E.[‡], & **Binz, M.**[‡,*] (2024). Human-like category learning by injecting ecological priors from large language models into neural networks. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 21121-21147).
   We propose a new framework for building computational models that are optimally adapted to naturalistic environments and show that such models capture human behavior across fifteen experiments from categorization, function learning, and decision-making.

9. Schubert, J. A., Jagadish, A. K., **Binz, M.**[‡,*], & Schulz, E.[‡] (2024). In-context learning agents are asymmetric belief updaters. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 43928-43946).

We demonstrate that in-context learning models exhibit asymmetric belief updating -- that is, they learn differently from positive versus negative prediction errors. This behavioral pattern closely mirrors what is commonly observed in human subjects.

10. Hussain, Z., **Binz, M.**[*], Mata, R., & Wulff, D. U. (2024). A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, *56*(8), 8214-8237.
This tutorial provides a practical and accessible guide to open-source language models tailored to behavioral scientists with limited technical background. Promoting the adoption of such models helps reduce reliance on proprietary tools and aligns with the ERC project's commitment to open science.

Legend: [†] co-first authorship, [‡] co-senior authorship, [*] publication without PhD advisor

PRIZES AND AWARDS

2022	German Cognitive Science Society Best Publication Award for best publication in cognitive science by a young investigator
2019	German Academic Exchange Service (DAAD) Scholarship for a three-month research visit at Harvard University
2019	EuroCogSci 2019 Best Poster Award for best poster presentation
2010	DMV-Abiturpreis for excellent performance in high school mathematics

SOCIETY MEMBERSHIPS

2024 –	European Laboratory for Learning and Intelligent Systems (ELLIS) Society
2024 –	Cognitive Science Society

INVITED TALKS

I have given over fifty invited talks, including keynotes, departmental colloquia, lab meetings, and workshop presentations, both nationally and internationally. A selection of these is listed below.

2025	Automated Scientific Discovery of Mind and Brain, Princeton, United States
2025	Japanese-American-German Frontiers of Science Symposium, Irvine, United States
2025	Inauguration of Excellence Cluster Reasonable Artificial Intelligence, Darmstadt, Germany
2025	Meeting on Relational Reasoning, Ghent, Belgium
2025	MIT Quest for Intelligence Seminar Series, Boston, United States
2025	Centre for Cognition, Computation and Modelling Seminar Series, Birkbeck, United Kingdom
2025	ELLIS Institute Scientific Symposium, Tübingen, Germany
2025	Max Planck Institute for Security and Privacy, Bochum, Germany
2025	Brown University, Providence, United States
2024	Google DeepMind, London, United Kingdom
2024	University of Oxford, Oxford, United Kingdom
2024	University of Milano-Bicocca, Milan, Italy
2024	Cognitive Sciences Colloquium, Irvine, United States
2023	nEuro-economics Seminar Series, Paris, France
2023	Cognition, Brain, & Behavior Research Seminar, Harvard, United States
2023	Colloquium of the Institute of Cognitive Science, Osnabrück, Germany
2023	International Titisee Conference on NeuroAI, Titisee, Germany
2022, 2023	International Interdisciplinary Computational Cognitive Science Summer School, Tübingen, Germany
2021, 2023	Reinforcement Learning and Decision-Making Seminar, Tübingen, Germany

ORGANISATION OF SCIENTIFIC MEETINGS

2025 Metacognition in Generative AI, EurIPS, Denmark
2025 Generative Adversarial Collaboration: Benchmarking in Cognitive Science, Conference on Cognitive Computational Neuroscience, Netherlands
2024 In-context Learning in Natural and Artificial Intelligence, The Annual Meeting of the Cognitive Science Society, Netherlands
2022 Meta-Learned Models of Cognition, The Biannual Conference of the German Cognitive Science Society, Germany

MAJOR INTERNATIONAL COLLABORATIONS

2019 – Prof. Samuel Gershman, Harvard University, United States
2020 – Dr. Eric Schulz, Helmholtz Munich, Germany
2021 – Prof. Matthew Botvinick, Google DeepMind, London
2021 – Dr. Jane Wang, Google DeepMind, London
2022 – Prof. Peter Dayan, Max Planck Institute for Biological Cybernetics, Germany
2023 – Prof. Thomas Griffiths, Princeton University, United States

ADDITIONAL INFORMATION

SUPERVISION ACTIVITIES

2023 – 2025 3 PhD and 1 Master students at the Institute for Human-Centered AI, Helmholtz Munich, Germany
2021 – 2023 2 PhD and 3 Master students at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany
2019 – 2021 2 Master students at the Department of Psychology, Philipps-Universität Marburg, Germany

TEACHING ACTIVITIES

2025 Guest Lecture, AI and Cognitive Science, Technical University of Munich
2022, 2023 Lecturer, Computational Cognitive Science, Eberhard Karls University of Tübingen
2020 Lecturer, Bayesian Statistics and Machine Learning, Philipps-Universität Marburg
2019, 2020 Lecturer, Theoretical Neuroscience, Philipps-Universität Marburg
2017 Teaching Assistant, Deep Learning, KTH Royal Institute of Technology

REVIEWING ACTIVITIES

2021 – Reviewer for Nature, Nature Communications, Nature Human Behaviour, PNAS, ICML, ICLR, NeurIPS, Psychological Review, Trends in Cognitive Sciences, Behavior Research Methods, Open Mind, Computational Brain & Behavior, Conference on Cognitive Computational Neuroscience, Annual Meeting of the Cognitive Science Society

OUTREACH ACTIVITIES

2025 Scientific Advisor for a book on AI and romantic relationships (in press)
2025 Press Coverage on “A foundation model to predict and capture human cognition” [New York Times, Frankfurter Allgemeine Zeitung, Tagesspiegel, MIT Technology Review, Nature, Science]
2023 Press Coverage on “Using cognitive psychology to understand GPT-3” [Science Journal for Kids, Tagesspiegel, Podcast]
2023 Presentation at Writing Competition on “Mensch und Machine” [Südwest Presse]
2019 Presentation at KFZ Science Slam, Marburg, Germany [YouTube]
2015 Mario Lives! Winner AAAI Video Competition [YouTube]

ERC Starting Grant 2026
Part B2

Part II of the Scientific Proposal (max. 7 pages, references do not count towards the page limits).

Theoretical integration remains one of the most pressing and unresolved challenges in cognitive science, contributing to what many now recognize as a widespread generalization crisis. To address this challenge, MIDAS will establish a new modeling framework that combines the predictive power of large language models with components grounded in cognitive science theory [1, 2, 3]. The project follows a tightly integrated methodology that brings together large-scale cognitive modeling, advanced machine learning techniques, and theory-driven interpretability. Its central objective is to develop the first large-scale, integrated model of human cognition. This will be achieved through three work packages:

- WP1 develops integrated models that are not only predictive, but also interpretable and grounded in cognitive science theory.
- WP2 establishes a robust methodological framework for evaluating and comparing such models at an unprecedented scale.
- WP3 pushes large-scale cognitive modeling beyond laboratory settings and into naturalistic, sensorimotor environments.

Each work package is designed to address a critical bottleneck in cognitive modeling, while building on the technical infrastructure and large-scale datasets established in my prior work. The following sections detail the technical implementation of each work package.

WP1: towards an integrated theory of cognition

Duration: M1-M48 • Personnel: PhD 1, Research assistant, PI

Developing an integrated theory of the human mind has long been a goal of cognitive science [4, 5]. Ideally, such a theory should manifest itself in a computational model that can *predict* and *explain* human behavior in any setting. My previous work took a significant step toward the *predictability* aspect of this goal [6]. The next step is to reintroduce cognition into the picture – moving beyond pure prediction toward models that offer *explanatory* insights into cognitive processes [7].

WP1 tackles this challenge by developing large-scale, integrated models that incorporate explicit, theory-driven assumptions about human cognition. Its research strategy follows the classical cognitive modeling pipeline: (1) constructing models with distinct architectural assumptions, (2) fitting them to large-scale behavioral data, and (3) evaluating their ability to capture human behavior. While grounded in established practices, WP1 scales this approach to a level not previously attempted – spanning many tasks (rather than just one) and using models with thousands of parameters (rather than a few). In doing so, WP1 will uncover both domain-specific and domain-general patterns of human cognition.

The transformer architecture – a highly flexible and empirically validated computational framework [8] – will serve as the starting point for model development. WP1 is organized into three tasks, which progressively introduce cognitive processes into transformers – implementing human-like mechanisms for memory (WP1a), learning (WP1b), and planning (WP1c).

Transformers are defined by their self-attention mechanism:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q} \in \mathbb{R}^{T \times D},\ \mathbf{K} = \mathbf{X}\mathbf{W}^{Q} \in \mathbb{R}^{T \times D},\ \mathbf{V} = \mathbf{X}\mathbf{W}^{V} \in \mathbb{R}^{T \times D}$$

$$\mathbf{Y} = \text{softmax}(\mathbf{Q}\mathbf{K}^{T})\mathbf{V} \quad (1)$$

This architecture offers a compelling entry point for integrative cognitive modeling. First, empirical results demonstrate that transformer-based models can accurately model human behavior across a wide range of cognitive domains [6, 9]. Moreover, their architecture is theoretically grounded, with rich connections to well-established models from neuroscience and cognitive science – including Hopfield networks [10], the Tolman-Eichenbaum machine [11], and the context maintenance and retrieval model of memory [12].

More specifically, MIDAS focuses on a variant of self-attention known as linear attention [13], which is obtained by removing the softmax operation from Equation 1:

$$\mathbf{Y} = (\mathbf{Q}\mathbf{K}^T)\mathbf{V} = \mathbf{Q}(\mathbf{K}^T\mathbf{V})$$

Importantly, linear attention can also be expressed in a recurrent form [14], enabling temporal processing that more closely reflects the sequential dynamics of human cognition:

$$\mathbf{W}_t = \mathbf{W}_{t-1} + \mathbf{V}_t \mathbf{K}_t^T \quad (2)$$

$$\mathbf{Y}_t = \mathbf{W}_t \mathbf{Q}_t$$

WP1a: memory

While effective for many tasks, the standard self-attention mechanism assumes a unified, unbounded memory in which all inputs are equally accessible – a sharp contrast to human memory, which is structured, capacity-limited, and governed by mechanisms such as forgetting and selective retrieval. To address this mismatch, WP1a embeds three core principles of human memory into Equation 2: (1) semantic memory, (2) forgetting, and (3) gating.

In particular, it will add semantic memory by introducing a learned, task-independent memory store $\mathbf{S}$ that encodes general knowledge about the world [15]:

$$\mathbf{W}_0 = \mathbf{S}$$

Forgetting will be implemented through a decay parameter λ applied to the memory traces [16]:

$$\mathbf{W}_t = \lambda\mathbf{W}_{t-1} + \mathbf{V}_t \mathbf{K}_t^T$$

Finally, gating will be introduced via a learned, context-sensitive controller $\mathbf{G}_t$ that modulates the flow of information into and out of memory [17]:

$$\mathbf{W}_t = \mathbf{G}_t \odot \mathbf{W}_{t-1} + \mathbf{V}_t \mathbf{K}_t^T$$

These components will be integrated into a unified architecture and tested empirically. More specifically, models will be fitted on Psych-101, a large-scale, cross-domain dataset of human behavior created in my previous work [6] and evaluated using a combination of held-out predictive accuracy and interpretability analyses. In particular, WP1a will examine how the different memory components operate over time (i.e., which mechanisms are engaged when) and how they generalize across tasks. Taken together, these innovations will establish a model architecture that is both predictive and explanatory, forming the foundation for integrative cognitive modeling in subsequent work packages.

WP1b: learning

While WP1a focuses on how information is stored and accessed, WP1b addresses a complementary question: how do people learn from experience? WP1b applies the same research strategy – theory-driven architectural design, parameter fitting, and empirical evaluation – to investigate the computational mechanisms of human learning.

Equation 2 implements a form of Hebbian learning [18], following the principle of "what fires together, wires together." In this formulation, the model updates its memory by strengthening associations between co-occurring units – enabling it to encode statistical regularities in the input without requiring explicit supervision.

However, Hebbian learning is just one part of the cognitive modeling toolbox. Error-driven learning, as formalized in approaches like the Rescorla-Wagner model [19], represents a fundamentally different approach. Instead of reinforcing co-occurrence, it adjusts internal states in proportion to the prediction error – the gap between what the system expects and what it observes. The linear attention framework allows for integration of error-driven learning, as shown below [14]:

$$\mathbf{W}_t = \mathbf{G}_t \odot \mathbf{W}_{t-1} + \alpha_t (\mathbf{V}_t - \mathbf{W}_{t-1} \mathbf{K}_t) \mathbf{K}_t^T$$

This formulation approximates gradient descent, taking a single step in the direction that minimizes the prediction error. Full optimization takes this idea one step further by directly computing the memory state that minimizes the prediction error $e(\mathbf{W}_{t-1})$, rather than approximating it through incremental updates [20]:

$$e(\mathbf{W}_{t-1}) = ||\mathbf{V}_t - \mathbf{W}_{t-1} \mathbf{K}_t||^2 + ||\mathbf{G}_t \odot (\mathbf{W}_t - \mathbf{W}_{t-1})||^2$$

WP1b systematically compares these three learning mechanisms: (1) Hebbian learning, (2) error-correcting learning, and (3) full optimization. Each approach embodies a distinct theoretical perspective on how humans adapt to experience. Like in WP1a, the models will be trained on Psych-101 and evaluated through a combination of held-out prediction and interpretability analyses, allowing to map out which learning strategies are deployed in which contexts.

WP1c: planning

WP1c investigates the role of planning in human cognition. In language-based models, the default approach to planning is chain-of-thought reasoning [21] – generating reasoning steps explicitly in natural language. However, this strategy is not well-suited to MIDAS's pipeline, as it is difficult to parallelize during model fitting and therefore computationally infeasible at the scale considered. WP1c instead explores an alternative technique known as latent planning [22], which is implemented by repeatedly applying the same layer to simulate multi-step reasoning within the model [23].

WP1c compares three variants of latent planning: (1) no planning, (2) planning with a fixed number of internal reasoning steps, and (3) planning with an adaptive number of steps [24]. Following the approach in WP1a and WP1b, models will be trained on Psych-101 and assessed using predictive accuracy and interpretability analyses, to identify which planning strategies align with human behavior across tasks.

Together, WP1 delivers an integrated modeling framework for memory, learning, and planning – three key pillars of human cognition. Each design choice is grounded in well-established theories and brings purely predictive models one step closer to capturing the cognitive processes of the human mind. **The outcome of this work package will be the first set of cognitive models that capture human behavior not just in isolated experiments, but across diverse domains – thereby marking a new era of cognitive modeling.**

WP2: methodological foundations for cognitive modeling at scale

Duration: M1-M36 • Personnel: Postdoc 1, Research assistant, PI

WP2 establishes the methodological backbone of MIDAS – developing scalable, principled tools for evaluating the large-scale cognitive models built in WP1. While traditional cognitive models are relatively small and easily evaluated, the models proposed in MIDAS operate at unprecedented scale and complexity. **WP2 develops the tools required to assess these models in ways that are scientifically meaningful, interpretable, and reliable.**

WP2a: evaluation metrics

The goal of WP1 is to develop integrated models that best characterize human behavior across a wide range of experiments. To identify the most plausible model from a set of candidates, one needs principled tools for model selection [25]. The standard approach in large-scale cognitive modeling is to use cross-validation: a model is trained on part of the data and evaluated on its ability to predict held-out samples [1, 2, 3]. However, what is ultimately needed are metrics that identify not just the model that predicts best, but the one that explains best. WP2a explores one such candidate: the model evidence, or marginal likelihood [26, 27].

The model evidence is the central quantity that arises when treating model selection as a problem of Bayesian inference. Bayesian model selection computes – for a given model M and some observed data D – a posterior distribution over models p(M|D). If one assumes a uniform prior over models, this posterior is proportional to the model evidence p(D|M). This quantity is obtained by marginalizing the likelihood over the prior distribution of the model's parameters:

$$p(D|M) = \int_{\boldsymbol{\theta}} p(D|M,\boldsymbol{\theta})p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta}$$

This approach is appealing because it rests on strong theoretical foundations and is widely used in conventional cognitive modeling, where models typically involve only a small number of parameters [25]. In such cases, exact or approximate computation of the model evidence is tractable and routinely applied. However, scaling these methods to the scale that MIDAS operates at has not been attempted so far as it involves major computational challenges.

WP2a explores how Bayesian model comparison can be applied to model selection for large-scale cognitive models. It builds on recent advances in deep learning, which have demonstrated that

approximate methods can be meaningfully used to compare models even in high-dimensional parameter spaces [28]. Laplace approximation, which estimates the model evidence by placing a Gaussian distribution around the maximum a posteriori estimate of the model's parameters $\boldsymbol{\theta}^*$, is of particular interest in this context:

$$\log p(D|M) \approx \log p(D,\boldsymbol{\theta}^*|M) - 0.5 \log |0.5\pi^{-1} \mathbf{H}_{\boldsymbol{\theta}^*}|$$

where $\mathbf{H}_{\boldsymbol{\theta}^*}$ denotes the negative Hessian at the maximum a posteriori estimate. This yields a tractable approximation of the marginal likelihood that balances model fit and complexity, effectively penalizing overfitting in accordance with Occam's razor [29]. I expect this approach to offer several concrete advantages: improved model recovery, stronger generalization to out-of-domain tasks, and a principled way to model individual differences. These benefits can be readily evaluated within WP1's pipeline, and the resulting insights directly integrated into the models developed there.

WP2b: model recovery

The central goal of MIDAS is to determine the cognitive mechanisms that give rise to human behavior. For this to succeed, one must ensure that our model selection procedures are trustworthy. Model recovery provides one safeguard for this [25]. The core idea is straightforward: simulate behavioral data from a known model M_1 and then evaluate whether a given inference procedure can correctly identify this model from a set of candidate models M_1, ..., M_n. In an ideal scenario, the data-generating model is reliably recovered, providing evidence that it is possible to distinguish between competing hypotheses. While model recovery is a standard tool for cognitive modeling, it has yet to be applied at the scale considered in MIDAS.

WP2b will take the models developed in WP1 and ask under what condition they can be reliably recovered. That will provide important insights into the structure of the model space (it should be constructed such that models are recoverable) and the selection of experiments (it should include experiments that facilitate recovery). WP2b will also assess the impact of different model selection metrics. In particular, it will compare the Bayesian approach developed in WP2a with more conventional selection procedures based on cross-validation. Together, these analyses close the loop between model development and evaluation, ensuring that MIDAS produces models that are both scientifically credible and computationally scalable.

WP2c: generative abilities

WP2c aims to develop a benchmark that measures the generative abilities of integrated cognitive models. While WP2a focuses on predictive accuracy – asking how well a model anticipates human behavior – WP2c addresses a similar but distinct question: do these models also *generate* human-like behavior? This distinction matters because accurate prediction does not guarantee that a model will reproduce the qualitative structure of human cognition [30]. Yet, generative ability is essential for many downstream applications such as in-silico prototyping [31].

WP2c consists of two components. The first aims to create a benchmark for evaluating generative behavior using a collection of well-established effects from the cognitive science literature, building on my earlier work on *CogBench* [32]. The current version of *CogBench* includes simulations for seven experiments testing ten behavioral effects. WP2c will expand this significantly by incorporating additional benchmarks from working memory [33], associative learning [34], decision-making [35], and others. Each experiment will be accompanied by statistical analyses to test whether a given model reproduces an effect previously identified in human behavior. The benchmark will score models on two dimensions: (1) the number of effects they capture, and (2) the similarity of their effect sizes to those observed in human participants. I expect that it is feasible to implement and evaluate 40-50 effects, thereby increasing the scale of *CogBench* by approximately five-fold.

The second component of WP2c goes beyond effects identified by human scientists and aims to develop a theory-agnostic method for benchmarking the generative abilities of integrated cognitive models. The goal is to evaluate whether a model can reproduce the structure of human behavior without being explicitly told what to look for, thereby avoiding confirmation bias. In essence, a truly human-like model should produce behavior that aligns with human data on any randomly chosen statistics. How can this be achieved? WP2c will use a randomly initialized neural network to encode two behavioral sequences – one from a human, one from a model – and quantify their

alignment by comparing the resulting internal representations. High similarity suggests that the sequences are behaviorally aligned. Importantly, this approach does not require the implementation of new simulators but can repurpose the ones already implemented. While this approach has proven effective in other domains – such as image [36] and language generation [37] – its application to human behavior is largely unexplored but has the potential to offer a scalable, unbiased metric for generative alignment.

WP3: integrated models for naturalistic environments

Duration: M13-M60 • Personnel: PhD 2, Research assistant, PI

The defining feature of the models developed in WP1 is that they operate on a domain-agnostic rather than a domain-specific representation – namely, natural language. This gives them a unique advantage: they can be applied not only to behavioral data from controlled laboratory experiments, but also to more naturalistic forms of human behavior, out-of-the-box and without architectural modifications [38]. **WP3 leverages this opportunity to extend cognitive modeling from controlled laboratory settings into the rich, complex environments where human cognition naturally unfolds.**

WP3a: human conversation

WP3a explores one of the richest forms of naturalistic behavior: human conversation. It examines whether the integrated models developed in WP1 can generalize to open-ended, real-world interaction. The result will be the first full-scale cognitive model of human conversation, made possible by the foundations laid in WP1 and WP2.

WP3a will take the integrated models developed in WP1 as a starting point but extend them in a critical direction: it asks how language interfaces with other cognitive processes [39]. There is a wealth of high-quality conversational data already available (e.g., from online forums [40], political debates [41], movie transcripts [42], chatbot interactions [43], and other sources) allowing WP3a to begin immediately without new data collection. The goals are twofold: (1) to determine how much one can infer about core cognitive processes purely from conversational data, and (2) to uncover how language interfaces with other cognitive processes such as memory, reasoning, and decision-making [39]. The results could demonstrate that language offers a uniquely rich signal for inferring cognitive processes – or clarify its limits in doing so. Either outcome would be foundational.

WP3b: high-dimensional environments

WP3b applies the same principles to high-dimensional sensorimotor data, asking how vision – rather than language – interacts with core cognitive processes [44]. For this, WP3b will augment the models developed in WP1 to incorporate visual input. The question of which visual representations best align with human perception has received considerable attention, but existing work has focused almost exclusively on static classification tasks [45, 46]. WP3b extends this line of inquiry to dynamic, agentic settings that require perception, action, and cognition to operate in tandem.

WP3b systematically compares three approaches to visual representation: (1) general-purpose, pretrained visual encoders (e.g., vision transformers trained on ImageNet [47]), (2) encoders trained end-to-end within the full model, and (3) models with integrated feedback mechanisms that allow for top-down modulation of perception by cognitive states [48]. Each of these offers different hypotheses about how the brain processes visual input: whether perception is largely feedforward and task-agnostic, optimized jointly with behavior, or shaped by higher-level cognitive goals.

The basis for building and testing these models will be a large-scale dataset of over 100,000 hours of human gameplay in a high-fidelity, real-time 3D environment that requires motor control, spatial reasoning, and long-term planning. This dataset has already been collected and is ready for use. WP3b thus sets the stage for a major leap in cognitive modeling, enabling the evaluation of models in environments that approximate real-world complexity more closely than ever before.

WP3c: the next-generation dataset

The long-term goal of MIDAS is to enable the creation of cognitive models for any computer-based tasks. This requires the right data, including recordings of human gameplay, expert applications, interactive tool use, and daily digital behaviors. However, no such multimodal dataset currently exists. WP3c asks: how should such a dataset look like and what would it take to build it?

The creation of such a dataset is beyond the scope of a single lab and will require coordinated, multi-institutional collaboration. WP3c will therefore organize a dedicated workshop bringing together leading researchers in cognitive science, machine learning, and human-computer interaction to develop the technical and logistical blueprint for this effort. The initial group of invitees will be drawn from my existing large-scale collaborations, including Psych-101 and Psych-201, and will be expanded as necessary to ensure full expertise and coverage. The workshop itself will address key questions around scope (what types of tasks and behaviors to include), infrastructure (data standards, collection protocols, annotation pipelines), and governance (access policies, ethical safeguards, and maintenance models). The resulting blueprint will lay the foundation for a future grant proposal – such as an ERC Synergy Grant – aimed at establishing the default dataset for modeling human cognition in the years to come. My prior experience leading large-scale collaborative projects, including Psych-101 (40 contributors) and the ongoing Psych-201 initiative (around 100 contributors), positions me perfectly to coordinate this next phase.

Helmholtz Munich postdoctoral incentive

Duration: M1-M60 • Personnel: Postdoc 2

Upon successful application, Helmholtz Munich will provide an additional postdoctoral position for the duration of the grant. This postdoc will focus on further advancing purely predictive models of cognition, complementing the main project's emphasis on explanatory modeling. More specifically, the postdoc will (1) lead the development of next-generation foundation models optimized for predicting human behavior; (2) scale up existing methods for leveraging such models for automated scientific discovery [49]; and (3) develop new approaches for extracting interpretable insights from them using mechanistic interpretability tools [50].

Work plan

Table 1: Work packages, tasks, timeline including the research team and their responsibilities.

	MIDAS	**Year**									
Work packages	**Tasks**	**1**		**2**		**3**		**4**		**5**	
Project coordination		PI	PI	PI	PI	PI	PI	PI	PI	PI	PI
WP1: towards an integrated theory of cognition	**A** memory	S1, RA	S1, RA	S1							
	B learning				S1	S1	S1				
	C planning							S1, RA	S1, RA		
WP2: methodological foundations for cognitive modeling at scale	**A** evaluation metrics	P1	P1								
	B model recovery			P1, RA	P1, RA						
	C generative abilities					P1, RA	P1, RA				
WP3: integrated models for naturalistic environments	**A** human conversation			S2	S2	S2					
	B high-dimensional environments						S2	S2	S2		
	C the next-generation dataset									S2, RA	S2, RA

Explanation regarding the research team: PI = Principal investigator, P1 = Postdoc 1, S1 = PhD Student 1, S2 = PhD Student 2, RA = Research assistant.

Table 1 provides an overview of the MIDAS timeline and project structure. My team consists of two postdoctoral researchers, two PhD students, and one research assistant, with backgrounds in psychology, cognitive science, machine learning, and computational modeling. Each core work package (WP1-WP3) will be led by a dedicated team member, while I will coordinate all aspects of the project and be directly involved in model development and theoretical integration. WP1 and WP2 will begin immediately and run in parallel to maximize synergies between model development and evaluation. WP3 will start with a one-year delay to build on the architectural and methodological insights produced by WP1 and WP2. The project aims to produce high-impact research, with target publications in top-tier interdisciplinary journals (e.g., Nature, Science, PNAS, Nature Human Behaviour) and leading machine learning conferences (e.g., NeurIPS, ICML, ICLR).

In line with my previous work, all data, code, and models will be released publicly upon completion, using established platforms such as GitHub, OSF, and Hugging Face to ensure maximum transparency and reusability.

Risk assessment and mitigating measures

MIDAS is an ambitious project, but it is built on a solid foundation of prior work, technical infrastructure, and domain expertise. My interdisciplinary background – at the intersection of cognitive science, machine learning, and language modeling – uniquely positions me to lead and execute this project. My earlier development of Centaur in particular demonstrates that I have the scientific and technical expertise required to build large-scale computational models [6]. Furthermore, the necessary computational resources are already in place: I have access to the Juwels Booster partition at Forschungszentrum Jülich – one of the most powerful GPU clusters in Europe – as well as a high-performance local cluster at Helmholtz Munich for rapid prototyping.

WP1 is a high-risk, high-reward effort. While it is already well established that large-scale transformer-based models can be trained to fit behavioral data effectively, the challenge lies in going beyond predictive accuracy to develop models that reflect the underlying cognitive processes. WP1 addresses this by integrating theory-informed structure into these architectures. Many of its components build directly on prior work, lending confidence that such models can be both performant and cognitively meaningful. The open question lies in the extent to which these models can reveal the mechanisms that underlie human thought. **Full success, however, would be truly ground-breaking.**

WP2 is designed to absorb some of the scientific risk inherent in WP1 by focusing on methodological development. Many of its tools – such as Bayesian model comparison – are well-established and have already been applied successfully in adjacent domains involving models of comparable scale. Their adaptation to large-scale cognitive modeling is technically feasible and represents a natural next step. Model recovery studies and evaluations of generative behavior are technically straightforward and rely on well-established procedures. Importantly, the success of the overall project does not depend on any specific outcome of these analyses – even partially negative results will yield important guidance for WP1. The main source of risk in WP2 lies in the design of a theory-agnostic method for conducting posterior predictive checks. While this is a relatively unexplored area in cognitive science, similar techniques have been developed in other domains, including image and language generation, which provide a solid foundation to build on.

WP3 is the most ambitious part of MIDAS, as it extends cognitive modeling into rich, naturalistic environments where no such models currently exist. There are early indications that human behavior can, in principle, be modeled in such settings, as demonstrated by proprietary systems [51, 52]. However, it remains an open question how the models developed in WP1 will transfer to these more dynamic tasks. Nevertheless, WP1 and WP2 provide a strong foundation for this effort, and even partial progress would yield valuable insights into how cognitive models can scale beyond the lab and into the real world. The risks in WP3 are mitigated by a staged progression in complexity, starting with conversational modeling, moving through modeling human gameplay within a single high-dimensional environment, and eventually culminating in the design of a large-scale behavioral dataset for naturalistic, computer-based tasks. This structure allows each component to build on the previous one while maintaining independent scientific value.

MIDAS directly tackles one of the central challenges in cognitive science: the **lack of theoretical integration** between different domains. It does so by introducing **the first large-scale, integrated models of human cognition** capable of simulating, predicting, and explaining behavior across a broad range of tasks and environments – something that does not yet exist in the field. These models will not only advance our scientific understanding of human cognition, but also **lay the foundations for future applications** in education, mental health, human-AI collaboration, and beyond.

References

[1] Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209-1214.
[2] Ji-An, L., Benna, M. K., & Mattar, M. G. (2025). Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 1-9.
[3] Eckstein, M., Summerfield, C., Daw, N., & Miller, K. J. (2024). Hybrid Neural-Cognitive Models Reveal How Memory Shapes Human Reward Learning.
[4] Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.
[5] Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.
[6] **Binz, M.**, Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... & Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, 1-8.
[7] Bowers, J., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. (2025). Centaur: A model without a theory.
[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
[9] Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., ... & Love, B. C. (2025). Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, *9*(2), 305-315.
[10] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., ... & Hochreiter, S. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.
[11] Whittington, J. C., Warren, J., & Behrens, T. E. (2021). Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*.
[12] Li, J. A., Zhou, C., Benna, M., & Mattar, M. G. (2024). Linking in-context learning in transformers to human episodic memory. *Advances in neural information processing systems*, *37*, 6180-6212.
[13] Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning* (pp. 5156-5165). PMLR.
[14] Schlag, I., Irie, K., & Schmidhuber, J. (2021). Linear transformers are secretly fast weight programmers. In *International conference on machine learning* (pp. 9355-9366). PMLR.
[15] Sukhbaatar, S., Grave, E., Lample, G., Jegou, H., & Joulin, A. (2019). Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*.
[16] Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., ... & Wei, F. (2023). Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.
[17] Yang, S., Wang, B., Shen, Y., Panda, R., & Kim, Y. (2023). Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*.
[18] Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology press.
[19] Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. *Inhibition and learning*, 301-336.
[20] von Oswald, J., Scherrer, N., Kobayashi, S., Versari, L., Yang, S., Schlegel, M., ... & Sacramento, J. (2025). MesaNet: Sequence Modeling by Locally Optimal Test-Time Training. *arXiv preprint arXiv:2506.05233*.
[21] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.
[22] Zhu, R. J., Peng, T., Cheng, T., Qu, X., Huang, J., Zhu, D., ... & Eshraghian, J. (2025). A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*.
[23] Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2018). Universal transformers. *arXiv preprint arXiv:1807.03819*.
[24] Graves, A. (2016). Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*.
[25] Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *elife*, *8*, e49547.
[26] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.
[27] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[28] Immer, A., Bauer, M., Fortuin, V., Rätsch, G., & Emtiyaz, K. M. (2021, July). Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning* (pp. 4563-4573). PMLR.
[29] Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic bulletin & review*, *4*(1), 79-95.
[30] Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, *21*(6), 425-433.
[31] Namazova, S., Brondetta, A., Strittmatter, Y., Nassar, M., & Musslick, S. (2025). Not Yet AlphaFold for the Mind: Evaluating Centaur as a Synthetic Participant. *arXiv preprint arXiv:2508.07887*.
[32] Coda-Forno, J., **Binz, M.**, Wang, J. X., & Schulz, E. CogBench: a large language model walks into a psychology lab. In *Forty-first International Conference on Machine Learning*.
[33] Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., ... & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological bulletin*, *144*(9), 885.
[34] Bhattasali, N. X., Tomov, M., & Gershman, S. (2021). CCNLab: A benchmarking framework for computational cognitive neuroscience. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
[35] Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., ... & Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature human behaviour*, *4*(6), 622-633.
[36] Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
[37] Chim, J., Ive, J., & Liakata, M. (2025). Evaluating synthetic data generation from user generated text. *Computational Linguistics*, *51*(1), 191-233.
[38] Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.
[39] Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289-312.
[40] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media* (Vol. 14, pp. 830-839).
[41] Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., ... & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852.
[42] Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.
[43] Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.
[44] Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*(3), 425-442.
[45] Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413-423.
[46] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.
[47] Demircan, C., Saanum, T., Pettini, L., **Binz, M.**, Baczkowski, B., Doeller, C., ... & Schulz, E. (2024). Evaluating alignment between humans and neural network representations in image-based learning tasks. *Advances in Neural Information Processing Systems*, *37*, 122406-122433.
[48] Noudoost, B., Chang, M. H., Steinmetz, N. A., & Moore, T. (2010). Top-down control of visual attention. *Current opinion in neurobiology*, *20*(2), 183-190.
[49] **Binz, M.**, Jagadish, A. K., Rmus, M., & Schulz, E. (2025). Automated scientific minimization of regret. *arXiv preprint arXiv:2505.17661*.

[50] Demircan, C., Saanum, T., Jagadish, A. K., **Binz, M.**, & Schulz, E. Sparse Autoencoders Reveal Temporal Difference Learning in Large Language Models. In *The Thirteenth International Conference on Learning Representations*.
[51] Team, A. A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., ... & Zhang, L. (2023). Human-timescale adaptation in an open-ended task space. *arXiv preprint arXiv:2301.07608*.
[52] Raad, M. A., Ahuja, A., Barros, C., Besse, F., Bolt, A., Bolton, A., ... & SIMA Team. (2024). Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*.

Appendix: All current grants and on-going / submitted grant applications of the PI (Funding ID)

Mandatory information (does not count towards page limits)

Current research grants (Please indicate "No funding" when applicable):

Project Title	*Funding source*	*Amount (Euros)*	*Period*	*Role of the PI*	*Relation to current ERC proposal*
No funding	-	-	-	-	-

On-going / submitted grant applications (Please indicate "None" when applicable):

Project Title	*Funding source*	*Amount (Euros)*	*Period*	*Role of the PI*	*Relation to current ERC proposal*
Unspoken rules: a data-driven approach to rule-based AI alignment by analysing human debate	AI Security Institute (AISI)	€206,166	2026	Co-PI	has no relation