Invited Symposia

INVITED SYMPOSIUM 1: NEURAL CORRELATES AND FUNCTIONS OF CONSCIOUSNESS IN HUMANS AND NONHUMAN ANIMALS

Organizer: Albert Newen (Ruhr-Universität Bochum)

Symposium Abstract:

Consciousness is a key phenomenon of our mental life but it still remains a riddle. How can we make progress in investigating consciousness?

In this symposium, we discuss some recent developments and perspectives including measurement of conscious experiences in nonhuman animals, especially birds (Nieder).

Furthermore, we argue that we need to strengthen an evolutionary and functional perspective to improve our insights. This includes conceptual distinctions of types of phenomenal consciousness based on evolutionary and functional evidence (Newen). Finally, we discuss further possibilities of accelerating research on consciousness given the state of art and new methodological approaches (Melloni).

Speakers

Lucia Melloni (Ruhr-Universität Bochum): Accelerating Research on Consciousness

Abstract: Thirty years ago, a seminal paper by Crick and Koch (re)introduced the scientific study of consciousness to the fields of psychology and neuroscience. This triggered a surge of research on the neural basis of consciousness, and concurrently the development of multiple empirically-based theories of consciousness. Since then, the science of consciousness has progressed from a nascent to a more established field. With growing maturity, new challenges emerge: how do we test the validity and predictive power of current theories? How do we know which theory best 'explains' consciousness? Consciousness research today is confronted with the question of how to move forward. How do we go from the accumulation of empirical findings supporting one theory or another, to solid theoretical foundations that can explain consciousness, predict its presence and absence, and improve the diagnosis and treatment of disorders in which consciousness is compromised? Here, we will discuss efforts underway to accelerate and transform research on consciousness based on best practices established in other fields, such as physics: large, collaborative team efforts oriented towards a common goal,

adversarial collaboration, preregistration, open data and open science protocols. The use of such practices has the potential to bring the science of consciousness forward and help arbitrate among competing theories. This approach entails a new sociology of science, and we will discuss how this by itself can help make progress, while also creating unique challenges. Given adequate resources and buy-in from our research community, these efforts may not only enable progress in research on consciousness but may also position our field at the frontier of science by providing a new model of how science can be done.

Andreas Nieder (Universität Tübingen): Neural correlates of consciousness in birds: A Bird's Eye View on Consciousness

Abstract: Determining whether animals possess subjective awareness of sensory stimuli, or sensory consciousness, remains a challenging qauestion. While consciousness is not necessarily required for complex behaviors, if present, it likely shares core features with human awareness. Working memory and voluntary attention, key components of human consciousness, are proposed as diagnostic markers for basic sensory awareness. Behavioral evidence suggests that these traits are present not only in mammals but also in birds, including corvid songbirds such as crows. In mammals, consciousness is closely tied to the cerebral cortex, with its unique layered structure and specialized cell types. However, recent neurophysiological studies in crows reveal a neuronal correlate of consciousness in their pallial endbrain. These pallial integration centers, which lack the cortical layering typical of mammals, arise from different embryonic pallial territories and feature independently evolved cell types. This challenges the notion that the mammalian cerebral cortex is a prerequisite for consciousness. Instead, it suggests that the anatomical and physiological properties of the telencephalic pallium provide a versatile substrate for the independent evolution of consciousness across vertebrate species, including birds. This highlights the possibility that different vertebrate lineages may evolve similar cognitive capabilities through divergent neural architectures.

Albert Newen (Ruhr-Universität Bochum): Types of phenomenal consciousness and their functional roles: unfolding the ALARM theory of consciousness

Abstract: The evolution of consciousness is a neglected topic that plays a surprisingly insignificant role in all major theories of consciousness. Furthermore, substantial disagreements can be observed in the dominant views on the neural correlates of consciousness, which focus too much on cortical brain regions. In order to dissolve some of the contradictions among these views, and in order to constrain the rival theories, we propose to distinguish three core phenomena of phenomenal consciousness: basic arousal, general alertness and reflexive (self-)consciousness. The central aim is to show that we can fruitfully distinguish specific functions for each of the three phenomena. Basic arousal has the function to alarm the body and secure survival by intervening in the slow updating of

homeostatic processes. General alertness fosters advanced learning and decision making processes, enabling various new behavioral strategies to deal with challenges; and reflexive (self-)consciousness enables future-directed long-term planning, accounting for the mindset of oneself and other agents. Constraining our contemporary theories of consciousness with this evolutionary and functional approach will enable the science of consciousness to make progress by accounting for three specific functions of consciousness, thereby informing the search for distinct neural correlates of consciousness (NCC).

INVITED SYMPOSIUM 2: SENSORY AUGMENTATION – EXTENDED MIND

Organizer: Julia Wolf (Ruhr-Universität Bochum)

Symposium Abstract:

TBA

Speakers

Silke Kärcher (feelspace GmbH): Augmented sensory experiences for spatial orientation: new tools based on magnetic field sensitivity to extend human cognition

> **Abstract:** Sensory augmentation offers a powerful avenue for exploring human perceptual plasticity. This work investigates how individuals learn to utilize novel sensory information delivered via wearable technology, focusing on spatial awareness and assistance for visually impaired individuals. Participants engaged with extended training using either a spatial navigation system (feelSpace) providing tactile north-directional feedback, or a hand navigation system (HANS) guiding grasping actions. Training with the spatial navigation system yielded significant behavioral benefits, including improved accuracy in survey and route knowledge within a virtual city, and increased use of spatial navigation strategies. These improvements appear to unfold through a three-stage model: initial activation requiring effortful processing and heightened awareness, followed by knowledge acquisition building detailed spatial representations, and culminating in deep integration where the augmented sense becomes seamless and automatic. Our research also demonstrated the efficacy of the hand navigation system (HANS) in assisting visually impaired users with grasping. The system successfully guided participants in complex tasks, navigating objects and obstacles in both controlled and real-world environments by translating visual inputs into actionable haptic feedback. These results highlight the potential of sensory

augmentation to enhance functional abilities and demonstrate the adaptability of the human sensory system, paving the way for technologies that enrich our perceptual experience.

Etienne Burdet (Imperial College London): Sensorimotor augmentation with humans and robots

Abstract: Sensorimotor augmentation could allow an individual to perform tasks that cannot be accomplished with just two arms and normal sensing. I will begin my talk with natural augmentation in polydactyl individuals born with six fingers on each hand, revealing how this grants them exceptional manipulation abilities. I will then review the state of the art in movement augmentation using supernumerary robotic limbs, where hardware prototypes have demonstrated feasibility, but intuitive and efficient control interfaces, effective sensory feedback, and a deeper understanding of the impact of augmentation on users remain open challenges. Finally, I will present how the human brain inconspicuously develops a model of interaction with the environment or with a human partner, enabling it to seamlessly integrate information gained from this partner in joint tasks.

Contributed Symposia

SYMPOSIUM 3: THE COMPLEXITY OF PAIN

Organizers: Sascha B. Fink (Friedrich-Alexander-Universität Erlangen-Nürnberg) & **Dominik Koesling** (Universität Münster)

Symposium Abstract:

Pain is a common and shared human experience. While nearly everyone encounters it, its complexity challenges our understanding. Pain varies in intensity, duration, location, association with bodily disturbances, and in how it is conceptualized, reported, expressed, and evaluated—reflecting its heterogeneity across individuals and cultures. Recent philosophical and scientific inquiries highlight this diversity, spurring interdisciplinary exploration across cognitive science, neuroscience, linguistics, and ethics. This symposium features four interconnected talks that examine pain's multifaceted nature.

Pain is increasingly understood not as a singular phenomenon but as encompassing diverse components. Neuroscience continues to struggle with identifying a universal neural mechanism for pain, leading to proposals such as the eliminativist approach, which questions the concept's utility (Corns, 2020). Linguistically, the simple term "pain" fails to capture its full complexity (Borg et al., 2019, 2020), often constraining understanding and stigmatizing

chronic pain patients. Pain's complexity is also influenced by cognitive and affective factors, particularly in chronic pain, where the interplay between emotional and cognitive mechanisms during different stages of chronification remains poorly understood. Metaphors and imagery further reveal pain's moral and cultural interpretations, shaping our understanding and potentially offering therapeutic benefits. Moreover, pain's complexity extends to memory, imagination, and empathy, where understanding its interaction with these processes depends on accurately ascertaining pain's multifaceted nature. This symposium seeks to deepen our understanding of pain in all its complexity and garner insights about its broader implications across disciplines and contexts.

Speakers

Sabrina Coninx (Vrije Universiteit Amsterdam), The complexities of pain – The problem of plenty and the problem of parts

> **Abstract:** Kripke (1981) posited that pain is defined by its "immediate phenomenological quality", contingently connected to C-fiber firing. Over decades, advancements in linguistics, philosophy of mind, and empirical research have challenged this view, bringing pain to the foreground of philosophical discourse. Still, its paradoxical and enigmatic nature fuels ongoing debate. My analysis of the complexities of pain begins with the assumption that pain episodes consist of at least four paradigmatic components: physiological disturbance, phenomenal experience, action tendencies, and mental network changes. This framework aids in exploring two key debates. First, the problem of parts: which components of pain are essential versus merely paradigmatic? This involves examining folk concepts, phenomenal characteristics of various pain episodes, and the existence of a unique pain quality. Second, the problem of plenty: how can the connections between the components enable us to naturalize pain? This involves analysis of the causation of pain as well as the neuroscientific foundation of pain in the light of its complexity and heterogeneity.

Frauke Nees (Christian Albrechts Universität Kiel): Beyond chronic pain: complex interplay between emotional and cognitive mechanisms.

> **Abstract:** In recent years, cognitive neuroscience has transformed our understanding of pain by exploring the neural mechanisms that shape pain experiences, from the synaptic pathways of peripheral nociceptors to the intricate networks linking brain regions involved in pain processing. However, the relationship between pain chronification and the emotional and cognitive mechanisms involved, as well as their adaptations at different stages of pain and their progression over time, remains poorly understood. This talk will address recent discussions on duration-dependent stages of pain, proposing that distinct mechanisms may underlie these stages. It highlights the risk of overlooking critical insights when pain mechanisms are examined solely through generalized chronic and subacute-chronic frameworks, which fail to account for

individual differences. Particular attention will be given to cognitive adaptations, such as coping strategies, in response to prolonged pain exposure. Recent findings in this area will be presented.

Claudia Bozzaro (Universität Münster): The ambivalent power of pain metaphors and images

Abstract: Pain is usually experienced and judged as something undesirable. Accordingly, pain relief is a generally accepted and undisputed moral duty and as a goal of modern medicine. This predominant negative view is also conveyed in various metaphors and images commonly used in relation to pain. While acute pain is repeatedly and almost unanimously metaphorically described as a warning signal, chronic pain is largely dominated by the image of an enemy and a battle' and is accordingly described as a foreign, disturbing evil, as a monster. This talk will address the significance of images in the context of pain in terms of three aspects: a) their epistemological function; b) their possible therapeutic/"healing" effect; c) their (ambivalent) normative power. The talk will not focus primarily on analyzing a specific image theory, but rather on reflecting on the normative implications of images, with a specific focus on normative implications of neuroimaging.

Ying-Tung Lin & **Christopher Jude McCarroll** (National Yang Ming Chiao Tung University): Remembering pain successfully

Abstract: Clinical pain assessment and management heavily rely on patient-reported symptoms and outcomes, which are frequently retrospective and depend on patients' successful recollection of pain. This reliance presupposes that certain aspects of pain are preserved in memory. This presentation explores the concept of successful pain memory at the intersection of the philosophy and science of pain and memory, as well as clinical practice. It explores which facets of pain—experiential components, affective components, associated physical events, action tendencies, temporal dimensions, etc.—must be captured in memory for recollection to be considered successful. Addressing these questions involves considering the nature of pain, as one's view on its constitution affects the perspective on pain memory. The issue is complicated by the heterogeneous nature of pain: The success condition for pain memory should not only account for paradigmatic cases of pain, such as acute pain but also the diversity of chronic pain.

SYMPOSIUM 4: AFFECTED BELIEFS – MECHANISMS UNDERLYING THE FORMATION AND CHANGE OF SELF-BELIEFS IN HUMANS AND AI

Organizers: Nele Rußwinkel (Universität zu Lübeck) & **Sören Krach** (Universität zu Lübeck)

Symposium Abstract:

Changes in human beliefs, especially in relation to oneself, occur only slowly (if at all) and are not always based on scientific evidence (e.g., the effects of human-induced climate change). However, while the focus of research to date has been on the when (e.g. in developmental psychology: when do self-beliefs develop?), the what (e.g., in general psychology: what is the consequence of a self belief for memory or attention processes?) and the why (e.g., in social psychology: why are people convinced that they belong to a group?), the question of how exactly people arrive at their self beliefs in the first place has received less attention. Current research shows that the formation of self-beliefs is closely linked to learning processes. However, the learning process, i.e. how and in what way information from the social environment is processed, is biased in a threefold sense and therefore not neutral: information processing is biased (i.e. 'affected') by 1.) the entire previous individual socialization history and thus pre-existing self-beliefs (i.e. 'priors'), 2.) by personal motivations (i.e. 'motives') and 3.) by accompanying emotional states (i.e. 'affective states'). In our symposium, we will bring together methods and concepts from experimental neuroscience (Krach), clinical psychology (Wilhelm-Groch) and computational modelling of behavior in the context of human-AI interaction (Russwinkel). We will discuss how people (but potentially also AI systems) may update their beliefs, anticipate belief updates in others and how these learning processes can be revised once they have been consolidated.

Speakers

Sören Krach (Klinik für Psychiatrie und Psychotherapie, Universität zu Lübeck): How are self-beliefs established and revisited?

> **Abstract**: Self-beliefs, such as beliefs about our abilities, attractiveness, or personality, are under constant (re)evaluation depending on the feedback we receive from our surrounding world 1. However, feedback processing is not a passive process during which information is picked up in an objective manner, rather the idea prevails that belief formation is essentially biased and shaped by affective and motivational processes. In several studies 2,3, we approach the question of how humans arrive at these self-beliefs in the first place 3,4 and, once established, how these self-beliefs are revised in the face of conflicting evidence 5. Using computational neuroscience methods, I will show that self belief formation is biased towards negative information and this bias is associated with the experience of affective states during belief formation. The findings support the notion that beliefs depend on global priors and are fundamentally shaped by motivational biases as well as affective experiences during feedback processing.
>
> References:

1. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. Nat. Neurosci. 14, 1475–1479 (2011).

2. Müller-Pinzler, L. et al. Neurocomputational mechanisms of affected beliefs. Commun Biol 5, 1241 (2022).

3. Müller-Pinzler, L. et al. Negativity-bias in forming beliefs about own abilities. Sci. Rep. 9, 14416 (2019).

4. Krach, S. et al. Examining self-belief formation through artificial beliefs. PsyArXiv (2024) doi:10.31234/osf.io/2y5tv.

5. Schröder, A. et al. Prior expectations about own abilities bias self-belief formation and hinder subsequent revision. bioRxiv 2024.08.30.610443 (2024) doi:10.1101/2024.08.30.610443.

Ines Wilhelm-Groch (Klinik für Psychiatrie und Psychotherapie, Universität zu Lübeck): Memory Reactivations effects on Negative Self-beliefs

Abstract: Negative self-beliefs (e.g. “I am inadequate”, I can’t trust that I will do the right thing”) play a significant role in the development and maintenance of various mental disorders such as post-traumatic stress disorder, depression, social anxiety and body dismorphic disorder. The modification of these self-beliefs is therefore a crucial mechanism of change in the psychotherapeutic treatment of these disorders. In my research group, we focus on negative self-beliefs that can emerge in the context of traumatic experiences. We aim to understand mediating role of trauma-related self-beliefs in treatment success and elucidate the potential of sleep-related interventions to target negative self-beliefs. In summary, our findings show that negative self-beliefs in traumatized individuals are often very stable and much harder to be modified than other trauma-related symptoms. Methods that affect memory formation during sleep such as targeted memory reactivation have the potential to accelerate the process of changing self-beliefs during psychotherapy.

Nele Rußwinkel (Institut für Informationssysteme, Universität zu Lübeck): The development of self-belief of being in control in human and artificial agents

Abstract: Beliefs about being in control of a situation can be affected by feedback on a sensory-motor control level as well on a cognitive control level. Different levels of expectations about the outcome of actions play a crucial role here. We have developed a conceptual framework (Kahl et al., 2022) how a sense of control develops and modelled a cognitive agent that can act in a dynamic environment with the mechanisms proposed. We compared the behaviour of the artificial and human agents

in different control situations. The influence of behaviour e.g., of visual attention and anticipatory behaviour can be well captured by the agent and can explain how humans are able to adapt to varying environmental changes and maintain a robust self-representation at the same time. These capabilities are crucial for artificial embodied agents that need to act flexibly in real world environments.

SYMPOSIUM 5: ANIMAL COGNITION FROM A COMPARATIVE PERSPECTIVE

Organizers: Maja Griem & Onur Güntürkün (Ruhr University Bochum)

Symposium Abstract:

Research in nonhuman animals is publishing astonishing insights into the mental life of nonhuman animals (animals, for short) with an increasing speed. This includes new unexpected species like bees and unexpected types of behavior, e.g. play behavior in wide variety of animals. In this symposium, we will present some key examples (Chittka and Griem) and then combine it with the methodological challenge of how to adequately investigate the mental life of animals. More precisely: How can we make progress in comparative cognition if we aim to systematically compare humans and nonhuman animals in principle without running into biased evaluations due to anthropomorphizing animal behavior, on the one hand, or overlooking animal competences due to lack of species-sensitive testing, on the other. Furthermore, we need to stop overintellectualizing human cognitive abilities ("anthropofabulation") if we develop a comparative perspective. We discuss some concrete cases of systematic progress in comparative cognition (Güntürkün) as well as propose a principle account, namely the multidimensional profile methodology for comparative cognition (Newen).

Speakers

Lars Chittka (Queen Mary University of London): Social insects - ancient civilisations?

> **Abstract:** The behavioural repertoires of social insects, their sophisticated social organisation and architectures and their foraging specialisations are unrivalled in the animal kingdom with the exception of the human species. Historically, however, these innovations have been regarded as "just innate" – as having been the result of evolutionary trial-and-error processes, with no element of learning, insight or culture. Recent work on the social learning capacities of bees call this simplistic view into question. Bumblebees do not just learn flower preferences from knowledgeable individuals – they can also learn object manipulation and puzzle-box opening techniques by observation. Some of their social learning feats even fulfil the basic criteria of cumulative culture, otherwise found only in primates. This makes it at least

cognitively plausible that some of the most remarkable behavioural accomplishments of social insects might, at least near their evolutionary roots, have been the results of individual innovation and subsequent social learning, and only later have become cemented into innate behaviours.

Maja Griem (Ruhr University Bochum): Are You Serious? Investigating Play Signals Yields Important Insights for Animal Communication

Abstract: Social play and the signals distinguishing play from serious contexts feature an important role in the development of humans and other animals. Play serves an outstanding role for the development of a broad variety of skills. Therefore, this paper focuses on the role of play signals for the establishment of social play contexts and for research in animal communication. Empirical studies on the function of play show that play serves as a multi-functional tool during the development of a variety of socio-cognitive skills not only in humans, but also in other mammals and possibly even other species living in complex social groups. Further, play signals are interesting cases of animal communication, as behavioral research supports the claim that play signals are used in an intentionally communicative way. In this paper I propose three criteria for intentional communication, namely (1) a learning component, (2) flexibility of use, and (3) a sensitivity to attentional states of others. Further, I argue that play signals can satisfy these criteria and therefore provide an exceptional opportunity to study communication across species. Further research on play signals may provide insights into the structure of social play, its intertwinement with socio-cognitive development, and animal communication.

Onur Güntürkün (Ruhr University Bochum): How to compare in Comparative Psychology

Abstract: In 1969, a groundbreaking work by Hodos and Campbell entitled “Scala naturae: why there is no theory in comparative psychology” shook the field of animal psychology. In fact, it makes no sense to compare animal cognition using a single test, as pure “success tests” along a single dimension obscure the deeper differences in species-specific cognitive strategies. But how can we compare animal cognition? And are such comparisons even meaningful? I will present three studies from my laboratory (conducted in collaboration with numerous colleagues from around the world) that address this question in different ways. The first study uses a series of cognitive tests that can be performed with different species to reveal different patterns of cognitive specialization. The second approach is a “signature test” that uses computer models to reconstruct how a task is solved by uncovering species-specific sensory-cognitive strategies and potential neuroevolutionary adaptations. The third approach is Q-

learning, which can be used to distinguish between learning rates and decision-making strategies. None of these approaches is perfect. But they may bring us closer to understanding the mind of another beast.

Albert Newen (Ruhr University Bochum): The multidimensional profile methodology (MPM) for comparative cognition

Abstract: How can we develop an adequate scientific understanding of the minds of nonhuman animals? We argue for a methodology based on multi-dimensional profile accounts. Such accounts are already used for the comparative study of norm cognition, consciousness, empathy and causal cognition, among others. This methodology demands that a cognitive capacity is characterized by a set of independent dimensions where each dimension is connected to operationalizable empirical indicators. Based on the level of realization for each indicator the level of implementation of a dimension is determined for a species, resulting in a multi-dimensional profile for each species. We analyze what this methodology is committed to. Then, we argue that this methodology has several benefits over competing unidimensional methodologies, by overcoming intractable disagreements, capturing the evolutionary continuity of cognition, alleviating anthropocentrism, and delivering more informative accounts of animal cognition. By demonstrating how this multidimensional methodology can be fruitfully combined with a methodology which focuses on the search for natural kinds in comparative cognition, we address the most important objection to the multidimensional profile methodology. We conclude that multidimensional profile accounts of all complex cognitive capacities should be developed and then used to facilitate scientific understanding of animal minds.

SYMPOSIUM 6: COGNITIVE ASPECTS OF TRUST IN HUMAN-AI TEAMS

Organizers: Ute Schmid (Universität Bamberg), **Eda Ismail-Tsaous** (Bayerisches Forschungsinstitut für Digitale Transformation) & **Celine Spannagl** (Bayerisches Forschungsinstitut für Digitale Transformation)

Symposium Abstract:

AI-based recommendation and classification systems, especially those relying on data-intensive machine learning methods, are becoming increasingly important in various application areas. Particularly in high-stakes domains such as medicine the aim is to ensure

both time-efficient and high-quality performance through the use of AI systems under human supervision. However, studies suggest that such hybrid human-AI teams do not necessarily perform better than humans or AI systems alone (Vaccaro et al., 2024). Rather, appropriate trust calibration is considered essential for team success, as overtrust can lead to unjustified over-delegation and a shift of responsibility, while undertrust can lead to negligence of correct system outputs. Explainable AI (XAI) methods have been proposed as a way to make AI systems more comprehensible and more trustworthy, but the relationship between explanations, trust and performance has proven to be complex, as various technical, psychological and ethical aspects need to be taken into account (Papenmeier et al., 2022; Longo et al., 2024).

Thus, understanding the conditions for successful human-AI teams demands a cooperative, interdisciplinary approach: The field of machine learning needs to address the development and implementation of faithful XAI methods (Longo et al., 2024). Cognitive psychology needs to uncover factors that influence user trust and allow for situation-based trust calibration, whereas philosophy has to provide ethical guidelines for the general design of human-AI interaction.

This symposium is part of the BMBF funded project “Ethical implications of hybrid teams of humans and artificial intelligence systems” (Ethyde), in which researchers from the fields of cognitive science, behavioral economics, philosophy and artificial intelligence work together.

References:

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106, 102301.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It’s Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. ACM Trans. Comput.-Hum. Interact. 29(4).

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat Hum Behav.

Speakers

Ute Schmid (Universität Bamberg): Explain to understand and explain to revise – Cognitive requirements for human-AI-teams

Abstract: With the growing number of applications for machine learned models there is an increasing demand for explainable AI (XAI) methods, which allow domain experts

to evaluate the trustworthiness of such AI-systems (Schmid, 2024; Thaler & Schmid, 2021). XAI approaches can be grouped into feature relevance methods, concept-based explanations, counterfactuals, and example- and prototype based explanations (Atzmüller et al., 2024). Furthermore, explanations are helpful in the context of human-in-the-loop learning where human feedback is used for model revision (Teso & Kersting, 2021; Slany, Scheele & Schmid, 2024). In this talk, I will present core concepts of explanatory interactive machine learning and point out open questions that should be explored empirically to gain insight into the effect of different types of explanations, as well as the possibility to correct such explanations, on calibrated trust and joint performance of human-AI teams.

References:

Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining. Data Mining and Knowledge Discovery, 38(5), 2571-2595.

Schmid, U. (2024). Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable. In: H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem (Eds.): Introduction to Digital Humanism (pp. 151-164). Springer.

Slany, E., Scheele, S., & Schmid, U. (2024). Hybrid Explanatory Interactive Machine Learning for Medical Diagnosis. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 105-116). Springer.

Teso, S., & Kersting, K. (2019). Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 239-245.

Thaler, A. M. & Schmid, U. (2021). Explaining machine learned relational concepts in visual domains effects of perceived accuracy on joint performance and trust. In Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 43, 1705-1711.

Johannes Fürnkranz (Johannes Kepler Universität Linz): Interpretability biases in machine and human learning

Abstract: Bias is a central concept in machine learning. It describes anything that is relevant for preferring one model over another, beyond the mere correctness on the training data. With the advent of powerful but intransparent machine learning models, the need for interpretable models has gained in importance. Yet, the question of interpretability of machine learning models is often reduced to a mere syntactic interpretability, i.e., to whether the model can be read and understood by a human or not. In this talk, we will argue that research in explainable AI should develop finer

grained distinctions between degrees of interpretability, and that human cognitive biases may be helpful to develop better XAI techniques. To understand interpretability, we must relate machine learning biases to cognitive biases, which let humans prefer certain explanations over others, even in cases when such a preference cannot be rationally justified. Only with such a collaborative effort can we develop suitable interpretability biases for machine learning.

References:

Fürnkranz, J., & Kliegr, T. (2018). The Need for Interpretability Biases. In: W. Duivesteijn, A. Siebes, A. Ukkonen (Eds.): Advances in Intelligent Data Analysis XVII. IDA 2018: 15-27.

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule based models. Mach. Learn. 109(4): 853-898.

Kliegr, T., Bahník, S., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artif. Intell. 295: 103458.

Fritz Becker (Universität Bielefeld): Perceived ability and competence as a factor of trust in human-AI teams

Abstract: How humans perceive expertise and trustworthiness in artificial agents depends heavily on their task expertise and domain knowledge. Observers who lack expertise rely on superficial cues, whereas experts can draw conclusions from the agent's actions (Lucassen & Schraagen, 2011). In this talk, I will present experimental results showing that novice observers struggle to accurately assess highly skilled agents in certain scenarios, often underestimating their competence. Conversely, expert observers evaluate agents more reliably, but still face limitations when agents surpass human level expertise. I will also present results on how an agent's reputation and performance affect the trust of human observers (Becker et al., 2024). Initial trust, influenced by external reputation, is recalibrated by observed agent performance. As an observer's perception of expertise is necessary to trust an agent (Mayer et al., 1995), these findings highlight the interplay between expertise perception and trust formation in human-agent interactions.

Fritz Becker, Celine Ina Spannagl, Jürgen Buder, Markus Huff, Performance rather than reputation affects humans' trust towards an artificial agent, Computers in Human Behavior: Artificial Humans, Volume 3, 2025.

Lucassen, T., & Schraagen, J. M. (2011). Factual accuracy and trust in information: The role of expertise. Journal of the American Society for Information Science and Technology, 62(7), 1232 1242.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709.

Sebastian Krügel (Universität Hohenheim): Decision-making context as a factor of trust in human-AI teams and associated ethical implications

Abstract: When employing AI-based decision support systems, there is a concern that trust in these systems may be misplaced. Academic research sometimes finds a form of over-reliance in these systems (e.g., Krügel, Ostermaier & Uhl, 2022, 2023a, 2023b) and sometimes a form of under-reliance (e.g., Dietvorst, Simmons & Massey, 2015, 2018). A priori, it is often unclear which form of mistrust will occur. Depending on the type of mistrust, however, different ethical issues arise, which ultimately also have an impact on developers at the level of the design of the AI-based decision support system. An important factor of trust in these systems appears to be the decision-making context. In this talk, I will discuss some empirical studies that find both over- and under-reliance in AI-based decision support systems and highlight the decision-making context as a possible explanation. I will also outline different ethical issues for both forms of trust.

Dietvorst, B.J., Simmons, J.P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144(1), 114–126.

Dietvorst, B.J., Simmons, J.P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science 64(3), 1155–1170.

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? humans trust untrustworthy AI advisors for ethical decisions. Philosophy & Technology 35(1), 17.

Krügel, S., Ostermaier, A., & Uhl, M. (2023a). Algorithms as partners in crime: A lesson in ethics by design. Computers in Human Behavior 138, 107483.

Krügel, S., Ostermaier, A., & Uhl, M. (2023b). ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports 13(1), 4569.

SYMPOSIUM 7: WHY CARE ABOUT CHATGPT'S THOUGHTS? – ON THE METHODOLOGICAL VALUE OF THE COGNITIVE SCIENCE OF AI

Organizer: Carlos Zednik (Technical University Eindhoven) & **Frank Jäkel** (Technische Universität Darmstadt)

Symposium Abstract:

Because of their high-dimensional, nonlinear and complex architecture, large language models (LLMs) are notoriously opaque. Because cognitive scientists are experts at explaining the behavior of high-dimensional nonlinear systems, they have been invited to engage in a "cognitive science of AI"—applying experimental techniques and modeling methods to explain these systems' behaviors (Rahwan et al. 2019; Taylor & Taylor 2021; Binz & Schulz 2023). Going beyond engineering concerns such as increased trustworthiness and usability, these cognitive scientists increasingly claim that explanations of language processing in LLMs might also provide insights into language-learning and knowledge-representation in humans (Mahowald et al., 2024; Zednik & Boelsen 2022).

How robust are these insights? Can they be justified? Some authors have argued that the scientific study of LLMs reveals "general principles" of cognition (Binz et al, 2024), either because the principles are domain general, or because the systems can be treated as "model organisms". Others highlight high-level similarities between LLMs and human brains and the possibility of "aligning" representations in artificial and biological neural networks (Mahowald et al., 2024; Merlin & Toneva 2024). In contrast, lingering doubts about the scientific value of such studies ground in LLMs' lack of biological and psychological plausibility, and in the fact that any insights gleaned about GPT_*n* might no longer apply to GPT_*n*+1. For this reason, although the "cognitive science of AI" may yield compelling explanations of machine behavior, it remains unclear whether, and if so to what extent, this can help us learn anything about human cognition.

This symposium brings together researchers from cognitive psychology, computer science, neuroscience, and philosophy of science to investigate the value, if any, that explanations of artificial intelligence have for the explanation of natural intelligence. In this way, it provides methodological guidance for the "cognitive science of AI", and situates this emerging discipline within cognitive science more broadly.

References:

Binz, M. & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. PNAS 120 (6): e2218523120. https://doi.org/10.1073/pnas.2218523120

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: 1–58. https://doi:10.1017/S0140525X23003266

Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). Machine behaviour. Nature 568: 477-486. https://doi.org/10.1038/s41586-019-1138-y

Mahowald, K., Ivanonva, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. Trends in Cognitive Sciences 28(6): 517-540. https://doi.org/10.1016/j.tics.2024.01.011

Merlin, G. & Toneva, M. (2022). Language models and brains align due to more than next-word prediction and word-level information. arXiv:2212.00596.

Taylor J.E.T. & Taylor G.W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. Psychonomic Bulletin & Review 28(2): 454-475. https://doi.org/10.3758/s13423-020-01825-5

Zednik, C. & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. Minds & Machines 32: 219–239. https://doi.org/10.1007/s11023-021 09583-

Speakers

Carlos Zednik (Technische Universität Eindhoven): Studying LLMs to explain human cognition – Context of discovery or context of justification?

> **Abstract:** Explainable AI is increasingly using methods from cognitive science to explain the behavior of artificial systems such as deep convolutional neural networks and large language models. It remains unclear whether, and if so how, this "cognitive science of AI" can also advance the science of natural intelligence. In this talk, I will attempt to clarify the relationship between XAI and cognitive science by applying Reichenbach's distinction between the context of discovery and the context of justification. Using this distinction, it will be possible to taxonomize and better evaluate the kinds of inferences that explanations of DCNN and LLM behavior permit about human and animal behavior. As the analysis will show, whereas many of these inferences yield pragmatic gains in the context of discovery, progress in the context of justification remains less secure.

Frank Jäkel (Technische Universität Darmstadt): On using natural language for self-programming in cognitive architectures

> **Abstract:** Humans can adapt their problem solving strategies. They can program their own behavior. They introspect, test, debug, and optimize their problem solving algorithms. These metacognitive activities can be implemented in standard cognitive

architectures that can store code in working memory and execute it with an interpreter that is implemented as a set of rules in a production system. Unfortunately, the programming language in which such mental code is written has remained elusive. I will argue that it is time to revive the old idea that this code is given in natural language. With the advent of large language models natural language interpreters might soon become an essential part of a new generation of cognitive architectures. In these architectures, the metacognitive activity of modifying your own programs might simply consist of transforming one natural language expression into another -- the task that transformers were developed for and are quite successful at.

Marcel Binz (Helmholtz Institut München): Foundation models of human cognition

Abstract: Most cognitive models are domain-specific, meaning that their scope is restricted to a single type of problem. The human mind, on the other hand, does not work like this – it is a unified system whose processes are deeply intertwined. In this talk, I will demonstrate how recent advances in large language models – together with a novel, large-scale data set that we recently collected – enable us to build unified computational models that predict and simulate human behavior in any experiment expressible in natural language. I will then show how methods from the mechanistic interpretability literature can be used to poke at the internal mechanisms of these models. This allows us to form new hypotheses about human information processing, ultimately leading to new insights on human cognition.

Polina Tsvilodub (Universität Tübingen): Scaling cognitive process models with scaffolded LLMs

Abstract: Computational modeling has been an instrumental tool for building explanatory models of human behavior in cognitive science (Farrell & Lewandowsky, 2018). However, current computational modeling is often limited along two dimensions. For instance, in the domain of human pragmatic language use, traditional cognitive models, while being explanatory, are often restricted to a prespecified closed set of contexts and utterances, while large neural network models, though applicable to more open-ended settings, lack explanatory transparency. In this talk, we present a hybrid approach, ScAffolded Generative models for Explanation (SAGE). SAGE leverages strengths from both approaches, using LLMs for scaling components that rely on processes grounded in open-ended world or language knowledge, scaffolding them within more explanatory process models. We present several models of language generation and interpretation in SAGE which successfully apply the framework to classical phenomena from pragmatics, while highlighting open questions and new directions for future research in computational cognitive science.

SYMPOSIUM 8: BRAINS IN SPACE - THE STRUCTURE AND METRIC OF THE COGNITIVE MAP

Organizer: Sen Cheng (Ruhr-Universität Bochum)

Symposium Abstract:

The term cognitive map first appeared in the late 1940s as a metaphor to account for the apparent ability of humans and animals to represent the spatial structure of their environment, drawing contrast to the prevailing behaviorist beliefs at the time. The later discovery of the presumed neural bases of this map in the form of place cells, grid cells, and other representations of spatial features led many to interpret this metaphor quite literally. To these researchers, the term cognitive map implies a globally consistent metric representation that encodes positions, distances, and angles in a Euclidean manner. The evidence for this view, however, is mixed, and the exact structure and metric of the cognitive map remain unclear. An alternative implementation of a cognitive map is a topological graph with nodes and edges representing connectivity rather than precise distances. However, neither of these alternatives alone is sufficient to explain all relevant experimental findings, perhaps suggesting some combination of the two. Furthermore, while the neural bases of spatial navigation have received much attention, with overwhelming evidence pointing to an essential role of the hippocampus and nearby structures, it has been difficult to pinpoint how the spatial representations in these regions might support the computations required for navigation using a cognitive map. In this symposium, the speakers will explore different views on the nature of the cognitive map from an interdisciplinary perspective that includes both experimental and theoretical work.

Speakers

Hanspeter A. Mallot (Universität Tübingen): Types of Spatial Representation: Reference Frames, Inner Structure, And a Note on Evolution

Abstract: The discussion of spatial memory and cognitive maps is complicated by the existence of multiple representations with different purpose and neural substrate. Sensory data and motor commands represent peripersonal space in an egocentric format. Path integration requires a representation of heading and allows the maintenance of an imagined viewpoint and a fixed imagined viewing direction. The result is an "allo-oriented chart" (Mallot 2023), i.e., a working memory reminiscent of a car navigation system in north-up mode. This chart is also used in route planning to determine the egocentric route decisions needed at each step. The referential memory of large-scale spaces does not change as the observer moves and is therefore "allocentric" in the sense of Klatzky (1998). It is organized as a graph of known places or local oriented charts that can be loaded into the working memory stage for

planning. The initial step in the evolution of (spatial) cognition is the representation of heading. It is based on similar mechanisms in both vertebrates and insects and thus seems to have evolved already in their last common ancestor, i.e., in early bilaterians some 600 million years ago.

Andrew Glennerster (University of Reading): A hierarchy of contexts for navigation

Abstract: Human memory is compositional: we learn broad categories and then subdivide these progressively to support more complex tasks. Navigational tasks are a good example. A policy (set of context-dependent actions) can describe movements based on a topological graph (which covers a broad category of potential paths) or, if the contexts are made more specific, a 'labelled graph', where lengths and angles of the edge-paths are more precise. An extreme version of this labelled graph could be behaviourally indistinguishable from a Euclidean representation. I will discuss this hierarchy in relation to published results from my lab and others measuring route-following and pointing performance in non-Euclidean mazes and link this to the hierarchical representation of 3D shape which progresses from crude disparity discrimination up to full Euclidean structure.

Sandhiya Vijayabaskaran (Ruhr-Universität Bochum): Emergent spatial representations in artificial agents

Abstract: The presence of allocentric spatial representations in the rodent hippocampal formation is often interpreted as evidence for the existence of a cognitive map. In this talk, I will present simulation results from a closed-loop reinforcement learning (RL) model that explains the emergence of such spatial representations. Our results suggest that both task demands and the use of an allocentric or egocentric reference frame affect the spatial representations that emerge in the agent. We also find that the choice of input representation affects behavior and spatial representations. Thus, the sensory inputs to the agent, the navigation task, and the output of the agent, i.e., every component of the closed loop, contribute to shaping the internal representations in the RL agent. These results will be discussed in the context of a hierarchical taxonomy of spatial navigation, with proposals of how the nature and metric of the cognitive map could be better understood through experimental manipulations and computational modeling.

William de Cothi (University College London): Predictive maps in and around the hippocampal formation

Abstract: The hippocampus, entorhinal cortex, and subiculum are known to encode an allocentric representation of spatial position, forming the neural basis of a cognitive map through diverse, spatially-selective receptive fields. The predictive map hypothesis provides a framework for understanding this cognitive map by modelling

spatial navigation as a reinforcement learning problem and factorising out transition and reward dynamics. By grounding these types of models in biologically plausible mechanisms, this talk will explore how predictive maps can account for a diverse range of observed phenomena in and around the hippocampal formation.

SYMPOSIUM 9: AUTOMATED SCIENTIFIC DISCOVERY OF MIND AND BRAIN

Organizers: Sebastian Musslick (Osnabrück University, Brown University), **Pascal Nieters** (Osnabrück University)

Symposium Abstract: Artificial intelligence (AI) is revolutionizing scientific discovery across disciplines, from physics to chemistry to biology—dramatically accelerating progress, as exemplified by AlphaFold's Nobel Prize-winning impact on the scientific discovery of proteins. Yet, cognitive science has only begun to explore the potential of AI-driven automation for understanding the human mind and brain [1]. This symposium introduces and discusses emerging approaches that leverage AI for automated scientific discovery in cognitive science. We highlight recent advances in automated model discovery, which enable the identification of novel mechanisms underlying neural computation and human cognition [2]; automated experimental design, which facilitates the discovery of new behavioral phenomena [3]; and automated research assistants, which integrate both capabilities in a closed-loop system to autonomously study human brain function and behavior [4]. Together, these innovations help extend the cognitive reach of human cognitive scientists, enabling them to explore larger spaces of experiments, models, and theories. However, these advances also pose critical challenges, from interpretability and reliability to the epistemological implications of AI-generated discoveries. This symposium will examine these challenges and outline future directions for harnessing AI to accelerate discoveries in cognitive science.

References:

[1] Sebastian Musslick et al. "Automating the practice of science: Opportunities, challenges, and implications". In: Proceedings of the National Academy of Sciences 122.5 (2025), e2401238121.

[2] David Weinhardt, Maria Eckstein, and Sebastian Musslick. "Computational discovery of human reinforcement learning dynamics from choice behavior." In: NeurIPS 2024 Workshop on Behavioral Machine Learning (2024).

[3] Sebastian Musslick, Younes Strittmatter, and Marina Dubova. "Closed-loop scientific discovery in the behavioral sciences." In: (Preprint) (2024). [Online]. Available at: https://doi.org/10.31234/osf.io/c2ytb.

[4] Sebastian Musslick et al. “AutoRA: Automated Research Assistant for Closed-Loop Empirical Research.” In: Journal of Open Source Software 9.104 (2024), 6839.

Speakers

Sebastian Musslick (Universität Osnabrück): Closed-loop scientific discovery in cognitive science

> **Abstract:** Closed-loop scientific discovery transforms empirical research by automating the cycle of data collection, modeling, and experimental design to generate new scientific knowledge. This talk introduces it as a paradigm for behavioral research in cognitive science, highlighting its potential and challenges. We formalize the research process as an iteration between data collection, computational inference, and experimental design, then present AutoRA, an open-source framework for automating behavioral research. We showcase its utility in discovering novel computational models and identifying new experiments. Specifically, we introduce automated model discovery methods that infer cognitive dynamics from human choice and reaction time data, yielding insights into reinforcement learning and decision-making. We also showcase automated experimental design methods for discovering novel behavioral phenomena in multitasking. The talk concludes by discussing challenges in cognitive science and future directions for refining closed-loop discovery systems and AI-scientists to advance our understanding of human cognition and behavior.

Daniel Weinhardt (Universität Osnabrück): Computational discovery of human reinforcement learning dynamics from choice Behavior

> **Abstract:** We present a novel machine learning framework designed to uncover human reinforcement learning models directly from choice data. By integrating recurrent neural networks (RNNs) with sparse identification of nonlinear dynamics (SINDy), our approach automates the identification of interpretable cognitive mechanisms underlying human decision-making. The method follows a two-step process: initially, an RNN is trained to predict choices of humans performing a reinforcement learning task, capturing their latent dynamics involved in learning. SINDy then extracts interpretable equations that represent the dynamical system learned by the RNN, revealing the cognitive mechanisms influencing participant behavior. Notably, this method allows for the exploration of structural differences in learning dynamics across individuals while capturing similarities across the group. We validated this framework on two-armed bandit tasks with synthetic and human data, where it consistently outperformed established models from human scientists. Our discovery method reveals novel cognitive mechanisms, providing fresh insights into human learning and decision-making.

Sedighe Raeisi (Universität Osnabrück): Computational discovery of individual differences in cognitive mechanisms

Abstract: Understanding individual differences in cognitive mechanisms across experimental paradigms remains a fundamental challenge in cognitive science. Traditional modeling approaches often assume a fixed model structure across participants, allowing only for variations in parameter values. However, cognitive processes may differ not only in parameter magnitudes but also in underlying structure across individuals. To address this limitation, we introduce a novel method that integrates Bayesian hierarchical inference with the sparse identification of nonlinear dynamical systems, enabling the discovery of distinct cognitive mechanisms within a population. We apply this method to a synthetic two-armed bandit task, modeling learning and forgetting dynamics across individuals. Our approach infers a distribution over model structures while inducing sparsity, allowing us to identify shared and individual-specific cognitive mechanisms. The results show that key statistical properties of the synthesized data are accurately recovered, demonstrating the method's promise in capturing structural differences in cognitive processes.

Pascal Nieters (Universität Osnabrück): *From neurons to cognition: Charting a data-driven path to dendritic computation*

Abstract: The neuron doctrine posits that the neuron is the fundamental unit of computation in the brain. However, advances in neurobiology reveal a more intricate picture, where dendrites exhibit compartmentalized, nonlinear properties that shape neural computation. While biophysically detailed simulations exist, computational neuroscience still relies on simplistic models like the leaky integrate-and-fire neuron. In contrast, emerging models of dendritic computation describe higher-level cognitive functions, such as symbolic reasoning, within single neurons. These models suggest that structural diversity among neurons is not a byproduct but a determinant of computational function. This raises a key challenge: how can we capture structural diversity across neurons from experimental data? Here, I propose a roadmap for leveraging data-driven computational discovery to address this gap, highlighting key challenges such as fitting bifurcating models to experimental data. This approach emphasizes the critical role of automated discovery techniques in achieving a comprehensive account of cognition grounded in neural data.