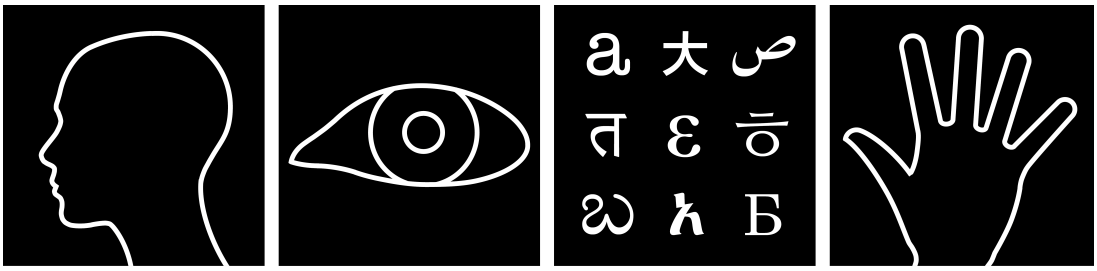

ESPP 2026 Book of Abstracts

33rd Annual Meeting of the
European Society for Philosophy and Psychology

30 June to 3 July 2026
Utrecht, the Netherlands

Keynotes and plenary symposia

Tuesday 30th June

09:00 Keynote. **Isabelle Dautriche**: The development of compositionality in language and thought

10:45 Plenary symposium with **Jean-Rémy Hochmann, Nina Kazanina & Iwan Williams**: *Format and structure of non-linguistic thought*

Wednesday 1st July

09:00 Keynote. **Mazviita Chirimuuta**: Evidencing Biological Naturalism

10:45 Plenary symposium with **Federico Adolfi, Dimitri Coelho Mollo & Beate Krickel**: *The Bases of Cognition: medium-(in)dependence, biological constraints, and the feasibility of computational-mechanistic explanation*

Thursday 2nd July

09:00 Keynote. **Ira Noveck**: Reconfiguring Figurative Language

10:45 Plenary symposium with **Nick Allott, Jane Dilkes & Diana Mazzarella**: *Understanding figurative language: functions, attitudes, and social meaning*

Friday 3rd July

09:00 Keynote. **Tadeusz Wiesław Zawidzki**: What Is Mindshaping?

10:45 Plenary symposium with **Ian Apperly, Víctor Fernández Castro & Ildikó Király**: *Mindshaping: Development, Diversity, and Individual Differences*

All abstracts

The abstracts of papers in this volume are organized according to the conference programme and are presented in chronological order. This structure allows readers to follow the sequence of sessions and presentations as they take place during the meeting.

Some programme entries do not have a separate abstract in this volume because the authors did not submit one. In those cases, please consult the original submission PDF, which is available via the conference website under “Programme and abstracts” (click on the title of the talk).

This volume does not include a table of contents. To locate a specific session, presentation, author, or title, please use your PDF reader’s search function.

The abstracts of the posters are at the end of the file.

Tuesday 30th June 09:00 — Keynote (Kerkzaal)

Isabelle Dautriche: *The development of compositionality in language and thought*

Compositionality is the property of a representational system whereby complex meanings arise from the combination of simpler meaningful elements. Here we take a developmental perspective to understand the origins of compositional representational systems and their relationship to language—an issue that bears on longstanding debates in cognitive science about the structure of the human mind. In this talk I will review some evidence suggesting that (i) infants begin to understand multiword combinations during the second year of life, and (ii) the key computation underlying compositionality—function application—can be observed in infancy during the first year of life, well before, and even outside, compositional language. These findings suggest the mind may support compositional operations prior to the emergence of compositional language, while leaving open questions about how the development of compositional thought and compositional language interact.

Tuesday 30th June 10:45 — Plenary Symposium (Kerkzaal)

Jean-Rémy Hochmann, Nina Kazanina & Iwan Williams: *Format and structure of non-linguistic thought*

Jean-Rémy Hochmann: *Representations of abstract relations in development*

Early in life, infants are capable of structured and sophisticated representations of scenes and events, articulating representations of entities (objects and agents) and their relations. However, the format of these representations remains a matter of debates. I contrast two competing accounts: discrete, language-like representations of relations and iconic, map-like representations. Drawing on a taxonomy of experimental tasks, I argue that discrete, language-like representations of relations emerge only with the acquisition of relational vocabulary, whereas younger infants rely on representations that are more accurately characterized as iconic and map-like.

Nina Kazanina: *Format and structure of (non-linguistic) thought: neurobiological foundations*

Since its appearance in 1975, Fodor’s Language of Thought (LoT) hypothesis has been influential in cognitive psychology and linguistics, but did not gain traction in cognitive neuroscience. The reason for the scepticism lies in the perception that neural implementation of a LoT is untenable. I disagree and demonstrate that critical ingredients needed for a neural

implementation of a LoT have in fact been found in rodents and other animals (Kazanina & Poeppel, 2023). I argue that cell types identified in animals in cognitive domains such as spatial navigation and numerical cognition instantiate exactly the representations and computations that the Language of Thought (LoT) framework calls for.

Iwan Williams: *Breaking the Language Barrier: Conceptual Representation without a Language-like Format*

An important part of the explanatory role of concepts is that they enable us to combine a wide variety of objects, properties and relations in thought, with contents spanning diverse domains. I discuss an argument that appears to show that paradigmatic non-linguistic representational formats are unsuited to play this role, and thus conceptual representation could not occur in these formats. I show that this argument fails, because it overlooks the possibility of individual concepts being shared between a number of special purpose representational systems. Demonstrating this requires defending the possibility of ross-format redeployment of concepts.

Parallel Sessions

Tuesday 30th June 14:30 — Symposium: New Directions in Episodic Future Thinking (Kerkzaal)

Episodic future thinking (EFT) is the ability to simulate oneself in a possible future scenario, which may relate to other future-oriented cognition. For the past 25 years, psychological research on EFT has mainly focused on lab-controlled tasks in which children or adults are told to describe themselves in a possible future event or choose a solution to a future problem. While providing us valuable information on the existence of a general future thinking mechanism, this research has only yielded results of how people use EFT in structured environments and vaguely indicates the type of mental representation an episodic future thought takes. Meanwhile, theories about the true nature of EFT and how it relates to other cognitive abilities have been sparse. We still do not know much about how people engage in EFT in naturalistic contexts (e.g., when somebody spontaneously packs extra socks because they know it will rain during their camping trip) or the individual differences that enable one to do so. Just as importantly, the role of EFT and how it interacts with other forms of imagination and future-oriented thinking have not been extensively considered.

Therefore, this symposium brings together philosophical and psychological perspectives to examine what EFT really is, how it operates spontaneously, how it interacts with other forms of imagination and temporal thought, and how it develops across the lifespan. Collectively, the talks challenge reductive views of EFT and highlight its role as a distinct form of future-oriented cognition that supports motivation, self-regulation, and planning.

The first talk opens by theoretically questioning the assumption that EFT can be reduced to imagining an event and placing it in the future. Instead, the speaker will introduce a non-reductive view of EFT, suggesting that simulating a future event is inseparable from understanding the relation between one's present and future self, offering a framework to understand emotional and motivational aspects attributed to EFT.

The second talk continues the philosophical thread by focusing on individual differences in future time perspective extension (FTPE) and how this impacts subsequent engagement in EFT. The speaker will address the history of future thinking research and its lack of consideration of an

individual's disposition towards the future, while also laying out the grounds for the future of psychological research in FTP regarding EFT.

The third talk will introduce a specific form of EFT– spontaneous EFT– and discuss the methodological challenges of conducting lab-controlled but naturalistic-oriented experiments to measure this ability in 5-7-year-old children. The speaker will discuss two experiments using iterative task refinements to demonstrate that children's spontaneous EFT is highly sensitive to motivational structure and constraints of the action space, highlighting the importance of task design.

The fourth talk will focus on implementing some of the previous speaker's guidelines, testing multi-step spontaneous EFT in 6-9-year-old children. The speaker will emphasize the importance of taking temporal ordering into account when measuring naturalistic EFT, and show that children can exhibit these naturalistic forms of temporally-specific EFT by 7 years of age.

Christoph Hoerl: *Episodic future thinking, imagination and time*

What is the relationship between episodic future thinking (EFT), on the one hand, and mere imagining, on the other? The equivalent question about episodic memory – i.e., how it relates to mere imagining – has dominated discussions about memory in both philosophy and psychology. Yet, when it comes to EFT, it is less clear that the question has been viewed as much of an issue. This may be because it is tempting to think that EFT simply consists in imagining an event plus placing it at some point in one's future, where these are two elements that can be described largely separately from each other. Call this a reductive view of EFT. I will discuss three reasons to look for an alternative to such a view of EFT: (i) studies examining potential benefits specifically of EFT (e.g., as opposed to other forms of imagination-involving or future oriented cognition) seem to implicitly rely on the assumption that the reductive view is incorrect; (ii) the reductive view arguably runs counter to parallels that have been drawn between EFT and other forms of 'self-projection', e.g., in the context of Theory of Mind abilities; (iii) the reductive view makes it difficult to explain the potential emotional consequences that EFT has been credited with. None of these considerations show conclusively that the reductive view is false, but they at least raise the question as to what an alternative, non-reductive, view of EFT might look like. I will offer a sketch of such a non-reductive view. This conceives of EFT as a form of perspective-taking in time, in which the aspect of simulating another perspective is not something that can happen separately from an understanding of the relationship between one's present perspective and the perspective one simulates.

Jenefer Husman: *Future Time Perspective Extension and Episodic Future Thinking: Toward an Integrated Framework*

Research on future thinking has consistently found that the further out into the future a goal is, the less impact that goal will have on a person's actions in the present. Activation of Episodic Future Thinking (EFT; the ability to pre-experience specific, vivid possible futures) and Future Time Perspective (FTP; an individual's perception and orientation toward the future) both mitigate the negative effects of temporal distance (Baird et al., 2021; Rösch et al., 2022). A few researchers have examined the intersection of EFT and FTP and have demonstrated that these aspects of future thinking are interrelated, and both influence participants' self-regulation (Gellert et al., 2011; Liu & Feng, 2019). This presentation will discuss an integrated framework that

demonstrates the interconnections between research on FTP and EFT, using a mixedmethods study to illustrate potential research directions.

Future Time Perspective has been one of the most researched aspects of human future thinking (Lewin, 1942; Nuttin, 1965; Lens et al., 2012; Husman et al., 2025). In the past two decades, cognitive neuroscience research on the neural similarities between memory and future thinking has opened another area of future thinking: Episodic Future Thinking (Rösch et al., 2022). Future Time Perspective has been examined as an individual difference trait, whereas EFT has been examined as a dynamic cognitive process. The intersection of these two factors has been examined, but the literature remains limited (e.g., Carr et al., 2021; Göllner et al., 2018).

Future Time Perspective is an individual difference variable. Each individual has a different time horizon: moments in time when goals that would be valuable in the short term become less valuable if they occur beyond that temporal boundary. FTP also encompasses the coherence or connectedness of goals within that temporal space. FTP can be conceptualized two dimensionally as an intersection of connectedness and extension. Individuals with high FTP will see connections between their current actions and future goals.

We argue that there is a relation between individuals' episodic future thinking and their Future Time Perspective. Specifically, FTP is related to an individual's ability to engage in EFT (Liu & Feng, 2019). In addition to presenting an integration of FTP and EFT research, we will report findings from a mixed-methods study examining longitudinal undergraduate students' FTPE and EFT. Drawing on semi-structured life-story interviews, students were asked to narrate key life chapters and to describe a detailed "day in the life" five years in the future. Narratives were analyzed using a multidimensional topographical framework of future time perspective extension (Spence et al., 2022) and coded for episodic future thinking features, including specificity and vividness. Quantitative measures of career commitment, future connectedness, and academic self-efficacy were used to contextualize these narratives. Longitudinal data allow us to examine both stability and change in future thinking within individuals over time, as well as differences across individuals at critical educational transition points. Together, these findings provide a concrete illustration of how FTP extension and episodic future thinking jointly shape motivation and self regulation.

Que Anh Pham: *Challenges of Developing a Spontaneous Episodic Future Thinking Task for Children*

Episodic future thinking (EFT) is the ability to self-project into the future (Atance & Meltzoff, 2005). Tasks studying EFT in children have investigated "cued" rather than "spontaneous" EFT – self-generating a solution to a future problem without environmental cues and verbal prompting. Spontaneous EFT is a better way of evaluating children's true EFT capacity (Atance et al., 2023); however, there are challenges associated with creating a spontaneous EFT task for children (Pham et al., 2024): How do children know that they have permission to interact with the environment? Will children want to interact with the environment without prompting? Will the lack of environmental cues cause confusion? To date, no published empirical task has successfully measured spontaneous EFT. Our task was developed with the goal of examining how to make EFT experiments more spontaneous while considering the aforementioned challenges.

In experiment 1, 60 children (20 5-year-olds, 20 6-year-olds, and 20 7-year-olds) visited the "Dinosaur Room," where they were shown that a drawing of a triangle made a sticker come out of a Sticker Machine. Then, the experimenter and the child went to the "Polka Dot Room," where they completed puzzles for 5 minutes. Subsequently, the experimenter gave the child a piece of paper and four differently colored crayons and said "You can do whatever you want now. After

this, we will go back to the Dinosaur Room." Through this design, we had created a mostly unprompted task with limited environmental cues. We found that 17.4% (SE=0.08) of 5-year-olds correctly drew the triangle, 18.2% (SE=0.08) of 6-year-olds did, and no 7-year-olds did. We suspected that children may not have had an intrinsic goal to retrieve a sticker without an explicit motivational factor and that the ability to draw freely may have been too overwhelming.

Therefore, we ran experiment 2 (ongoing) where we sought to increase children's motivation by offering them a toy from a treasure chest if they had enough stickers. Furthermore, we constrained the possible drawing space in the Polka Dot Room, where we only gave children one black crayon instead of four colored crayons. Our rationale was that the possible drawing space was too expansive in experiment 1 and that by limiting the drawing utensil to a neutral color, children would be less biased to draw things in their everyday lives. Currently, 7 out of 7 5-year-olds and 2 out of 6 6-year-olds have drawn the correct shape.

Ultimately, results from experiment 2 will allow us to be more certain about the factors impeding children's spontaneous EFT performance, notably the lack of direction (i.e., from a lack of explicit prompting) causing children to be overwhelmed by the possible actions (i.e., drawing anything they wanted). By the conclusion of the study, we will have a deeper understanding of how to narrow the space of possibilities and increase motivation and goal-directed behavior in a spontaneous EFT task for children.

Jade Zack: *The development of temporal ordering for spontaneous episodic future thinking when solving a multi-step future problem in middle childhood*

Spontaneous episodic future thinking (EFT) is the mental projection of oneself into a future scenario without being explicitly prompted (Atance et al., 2023; Pham et al., 2024). In real life, people often engage in multi-step spontaneous EFT: they simulate the individual steps needed to achieve a future goal and the specific temporal order to take these steps in (for instance, simulating first finding a key that opens a closet and then taking out a coat for an upcoming snowstorm; Martin-Ordas, 2018). However, very few studies have examined the process of completing multiple steps during EFT in children, and none have tested it without prompts. As such, this talk will focus on the importance of taking temporal ordering into account when measuring naturalistic future thinking by highlighting a study that measured children's use of spontaneous EFT to solve a multi-step future problem.

In our computerized task, children had to open a magic box by finding target words and saying them aloud in a specific order. First, children were introduced to the general procedure of choosing objects for the future ("travel phase"; Pham et al., 2024). Next, in a training phase, they learned how to open a box by seeking out "magic" words and saying them in a specific order to receive a prize. Finally, in the test phase, children learned that they needed to find words according to a new rule to open the box. After a delay, they (virtually) went to another room with target and distractor words, as well as distractor activities, and were allowed to act freely. After choosing words, children went back to the first room and were given the chance to open the box. Children's choice of words (target vs. distractor) measured their spontaneous EFT performance, and the order they said the words served as a measure of their temporal order performance.

Currently, 114 children, 27 6-year-olds, 43 7-year-olds, and 44 8-year-olds have participated in this study (final N=138). We found that children's spontaneous EFT performance improved with age ($p= 0.035$, OR=1.043, 95% CI [1.004, 1.09]). Follow-up binomial tests showed that while 6-year-olds were at chance level (9/27 = 33%), 7-year-olds (22/42= 51%) and 8-year-olds (27/44= 61%) chose the correct words significantly above chance without prompts ($p < .01$). Non prompted

temporal order performance, however, remained significantly above chance levels (M = 88%), regardless of age (OR = 1.03, 95% CI [0.96, 1.10]) or spontaneous EFT performance (OR = 1.44, 95% CI [0.39, 5.19]).

As children age into independence, we must understand the intricacies of their realworld planning processes, and this study is a step towards this. Our findings suggest that naturalistic planning (multi-step spontaneous EFT) begins to develop around 6-7 years old and continues to improve with age. These results also suggest that temporal ordering is embedded in– and possibly inseparable from– multi-step future thinking, and therefore, future EFT research must properly take it into account.

Tuesday 30th June 14:30 — Vehicles of Representation (Grote zolder)

Matteo Colombo: *The Medium-Independence of Computation, Neural Computing, and Neurotransmitters*

A common view in philosophy of mind and cognitive science is that, if the brain computes, then all its computations must be medium-independent procedures. This is because all physical computations are essentially medium-independent, since computation is an abstract, formally defined procedure, which can be physically implemented (or realized) in many different media. It follows that all neural computations must be medium-independent (Anderson & Piccinini 2024; Drayson 2025; Garson 2003; Haugeland 1985; Piccinini & Shagrir 2014; Piccinini 2020; Piccinini & Bahar 2012).

The operations performed by my gastrointestinal tract to digest what I eat are instead medium-dependent. This is because the operations constitutive of digestion "are defined in terms of specific physical alterations of specific substances," such as specific chemical reactions involving specific molecules (Piccinini 2015, 122; Shagrir 2022, Ch. 5). So, digestion, unlike computation, cannot be physically implemented in many physical media, and so, it is not a medium-independent procedure.

Digital computation is the paradigmatic example of medium-independent computation, whereby "[t]he rules defining digital computations are defined in terms of strings of digits and internal states of the system, which are simply states that the physical system can distinguish from one another. No further physical properties of a physical medium are relevant to whether they implement digital computations. Thus, digital computations can be implemented by any physical medium with the right degrees of freedom" (Piccinini 2015, 123).

Candidate vehicles of computation in the brain—what computational neuroscientists call the neural code (Rieke et al. 1997; deCharms & Zador 2000; Brette 2015; Cao 2018)—include the rate and timing of neural action potentials (or spikes). These would be medium-independent vehicles for neural computation, because "[t]he functionally relevant aspects of spike trains, such as spike rates and spike timing, are similar throughout the nervous system regardless of the physical properties of the stimuli (i.e., auditory, visual, and somatosensory) and may be implemented either by neural tissue or by some other physical medium, such as a silicon-based circuit" (Piccinini & Bahar 2013, 462). The problem with this suggestion is that we do not know whether all and only properties that define the neural code are properties of action potentials.

The nervous system has electrical properties. But it also possesses biochemical, molecular, morphological, and historical properties (Striedter 2005; Sterling & Laughlin 2015; Sporns 2016; Zeng & Sanes 2017; Cao 2022; Chirimuuta 2022). Although we do know that neural transmission can be electrical or chemical (Pereda 2014; Valenstein 2005), nobody has yet demonstrated that neurotransmitters cannot be vehicles of computation in the brain. If certain neurotransmitters

constitute vehicles of neural computation or are directly involved in specific neurocognitive operations, then it will follow that some neurocognitive operations are not neural computations, which narrows the scope of the computational theory of mind. Or, it follows that the brain does not literally implement the sort of computations formalized by existing mathematical theories, which supports the idea that the brain should not be understood as a computing system. Or, it follows that not all neural computations are essentially medium-independent, which calls into question the fruitfulness of the notion of medium-independence in computational neuroscience.

My aim in this paper is to provide evidence that certain neurotransmitters do constitute, at least partly, vehicles of computation in the brain.

Andreas Pommer: *Population-Level Representations and Computations Afford Causal Explanations*

The neural population doctrine has emerged as a central framework for explaining cognitive capacities with the advent of large-scale neural recording techniques (Barack & Krakauer, 2021; Ebitz & Hayden, 2021; Yuste, 2015). A prominent approach models population activity in a state space, where low-dimensional patterns of activity called manifolds represent the task-relevant variables and the computations are performed by the dynamics of the population state over the manifold (Vyas et al., 2020). Despite the prominent role of these models in explaining cognitive capacities, their explanatory status has received little philosophical attention. Because of the high-level and mathematical nature of these models, they may appear as mere phenomenological models that only compactly describe the phenomenon.

In this talk, I argue to the contrary that they afford proper explanations by capturing difference-makers at the population-level. Thus, I show how representational vehicles and computations specified at the level of the population may provide causal explanations of cognitive capacities. First, *I argue for an interventionist approach to support population-level representational vehicles and computations as difference-makers*. A common way to demonstrate that a model is explanatory is to provide a mapping onto the parts of the system that play causal roles. To this end, a popular approach relies on decomposition and localisation (Bechtel & Richardson, 1993) and has been used to show how dynamical models can be explanatory (Bechtel & Abrahamsen, 2010; Kaplan & Craver, 2011). I argue that this approach fails to accommodate the representational vehicles which are specified at the population-level due to the mixed selectivity of many cortical neurons (Fusi et al., 2016; Hardcastle et al., 2017; Shea, 2007). Population-level vehicles have been identified in previous work on connectionist and neural systems (Burnston, 2021; O'Brien & Opie, 2006; Shagrir, 2012), and a key challenge is to account for how population-level vehicles may play causal roles and figure in genuine representational explanations (Ramsey, 2007). This is precluded if computations are taken to occur at the level of individual neurons.

To resolve this, I apply the interventionist account of causation (Woodward, 2003), which provides level-neutral criteria for assessing whether population-level variables are difference-makers for cognitive capacities. Next, *I argue that dynamical models of population activity specify population-level difference-makers for explanations of cognitive capacities*. Analysing two examples from the empirical literature (Mante et al., 2013; Sohn et al., 2019), I argue that the dynamical models satisfy the interventionist criteria for causal explanation. Specifically, they provide invariant counterfactual dependencies between the computation performed by the network and interventions on both the input and the population-level variables. Moreover, I argue that these population-level variables are genuine representational vehicles because we can

intervene on their particular content such that a counterfactual relation additionally exist between the content of the vehicles and the output of the system (Ramsey, 2007; Schulte, 2023).

I conclude, therefore, that the dynamical models identify population-level difference-makers that support answers to what-if-things-had-been-different questions that figure in genuine representational explanations. A worry that may occur is that because the variables—the explanans—are population-level properties then they can only serve to redescribe but not explain the computation, i.e., the explanandum (Chirimuuta, 2018). Against this interpretation, I argue that the explanans enable surgical and independent interventions that can be used to change specific properties of the network computations, and, moreover, that we can intervene on the computation of the network to change the explanans of the models (Sadtler et al., 2014). This is akin to the mutual manipulability criteria developed within the mechanistic framework (Craver, 2007). Consequently, I conclude that the dynamical models provide explanations of how the computations are produced on an interventionist account of causal explanation.

Finally, *I respond to objections that the models fail to satisfy the interventionist requirements for causal explanation*. One might object that the models fail to provide counterfactual dependencies that can be interpreted as outcomes of interventions. If one accepts that the models answer w-questions, these models would then provide non-causal explanations (Chirimuuta, 2018). The real causal work—the computations—might instead be placed at the level of individual neurons. In response to these objections, I first argue that the population-level variables constitute the appropriate level of explanation before demonstrating that they satisfy the criteria for intervention. An important part in developing causal explanations is to find stable difference-makers that are proportional to the explanandum. I argue those considerations favour the low-dimensional manifolds and dynamics provided by dynamical models (Woodward, 2021). First, these models are stable across changes in background conditions, including behavioural states and animals (e.g., Chaudhuri et al., 2019; Nieh et al., 2021). Second, the population-level variables are invariant across changes in the neural population and capture the appropriate contrastive focus with the explanandum (Gallego et al., 2020; Woodward, 2008). There are therefore good reasons to take seriously the population-level variables as difference-makers. In evaluating whether they are targets for interventions, I first consider what variables need to be held fixed when assessing causal claims (Shapiro & Sober, 2007; Woodward, 2015).

On this basis, I argue that the causal powers of the population-level variables are not excluded by their underlying realisation base. Moreover, I argue that since the variables can be intervened on independently to change the explanandum while keeping other variables fixed, the criteria for unconfounded interventions are satisfied (cf. Woodward, 2025). Second, I argue that the information required to answer w-questions is available exclusively at the population-level and can be exploited with precise methods in experimental practice (Vinograd et al., 2024). I therefore conclude that dynamical models of population activity satisfy the interventionist criteria for causal explanations and thus explain cognitive capacities through population-level difference-makers.

Halely Balaban, Shachar Lando & Roy Luria: *Dissociating pointers and representational content in visual working memory*

Representation is a key concept across all fields of cognitive science, both highly influential and admittedly elusive. In psychology, one important representational mechanism is working memory (WM), an online workspace maintaining a limited amount of relevant information in an active state, ready to be accessed and manipulated (Baddeley, 1992). WM's representational contents tightly

affect our internal thoughts and external actions, and accordingly they have been, and still are, the focus of much research.

A growing body of theoretical and empirical work has suggested that to permit their online status, WM representations must be mapped to the external environment via a pointer system (Pylyshyn, 2000). Each pointer is a one-to-one mapping, through which the correct WM-representation can be accessed and updated if the corresponding real-world object changes. However, different frameworks diverge regarding the specific relationship between WM representations and pointers. Several prominent current views (e.g., Awh & Vogel, 2025) equate pointers with WM representational units (i.e., slots), resonating the object files (Kahneman et al., 1992) idea of episodic bindings. Yet, a different conception of WM-pointers explicitly distinguishes them from the level of representations; in this view (e.g., Balaban & Luria, 2025), pointers only index WM representational-units but cannot be equated with them, because they do not directly represent anything about objects besides their individuation. The unique prediction of this approach is that it should be possible to create situations of a mismatch between the number of representational-units and the number of pointers in WM, an idea that clearly contradicts the claim that pointers equal WM-representational units.

To examine the dynamics of representations and pointers in WM, here we rely on a thoroughly validated event-related potential (ERP) component named the contralateral delay activity (CDA; Vogel & Machizawa, 2004). The CDA amplitude rises – in a highly selective manner (e.g., dissociable from perceptual or attentional factors; Luria et al., 2016) – with representational load in visual WM. This hallmark set-size effect means the CDA indexes how many representational-units are held in WM across conditions or timepoints. Recently, it was further shown that events that disrupt the ongoing function of WM's pointer system cause a transient decrease in CDA amplitude, in line with a 'resetting' process, whereby unmapped representations are removed and replaced by new (mapped) ones (Balaban & Luria, 2017). This happens, for example, when a coherent object splits in two, presumably because pre-split a single pointer supported the representation, and post-split neither of the two objects matches the original representation. The CDA-drop flags situations that invalidate the WM-environment mapping, as opposed to almost identical situations that are updatable via a stable mapping (e.g., when each half is clearly marked already pre-split), producing monotonic changes in CDA amplitude. As such, the CDA-drop serves as an index of pointer status that can also uncover the number of active pointers.

Can the two levels of WM (representations vs. pointers), as indexed by the two aspects of the CDA (set-size effect vs. CDA-drop), be misaligned? Here, we focus on WM information-compression, where the representations of several things (features, objects, etc.) are chunked and held in a single representational-unit in WM. While strong integration (e.g., the different parts of a onesie) should rely on a single 'fused' pointer, we argue that sometimes the pointers can remain independent despite the representations being compressed. Specifically, when several distinct objects are grouped into one representational-unit (e.g., the shirt and pants that make an outfit), we hypothesizes that each represented object within the group maintains its own mapping with the corresponding environmental object. Across five EEG experiments (N=80 total), dissociations between the CDA set-size effect and the CDA-drop supported this claim.

The first study contrasted strong and weak integration within the same task, using highly similar stimuli. Participants performed a shape change detection task on items that could either come together, break apart, or move in a steady manner throughout the trial. The movement was task-irrelevant, and served only to dynamically manipulate integration cues. In the critical condition, two shape-halves met to form a composite shape that moved as a single whole, and then separated again. The joint movement period fostered integration, as reflected by decreased CDA amplitudes, both when the two shape-halves were uniformly black and when each had a unique task-irrelevant color. But the later split resulted in markedly different patterns: an updating process

with a steady amplitude rise for uniquely-colored halves that could be easily individuated, and a resetting process with a CDA-drop for uniformly-colored halves. So, situations that are similarly compressed in terms of representational-units can be supported by a varying number of pointers depending on the strength of integration; grouping distinct objects is supported by multiple pointers, while unifying the parts of a single object involves only a single pointer.

The second study focused on perceptual grouping, and independently manipulated the coherence of either the group or the constituting objects. Participants performed an orientation change detection task with three Pacman shapes that either moved independently or formed a Kanizsa triangle (motion again being task-irrelevant), whose lower CDA amplitude indicated successful integration. When one or all of the Pacman items left the group, the CDA steadily rose, indicating updating into partial or full (respectively) individuation. But when only half a Pacman left the group, a CDA-drop indicated resetting, despite the group structure remaining largely intact. Thus, disrupting the group structure while preserving the items did not influence the pointers, suggesting no group-specific pointer. In contrast, disrupting an item while preserving the group invalidated the pointers, demonstrating that grouped objects retain their individual pointers despite being compressed into a single WM representation.

Together, the results show that WM-grouping involves compressing representations without affecting the pointer system. The still-independent pointers make grouping flexible, supporting unpacking/repacking as needed. More generally, the findings dissociate between pointers and representations in visual WM, highlighting the importance of treating these two levels as distinct. Pointers form the infrastructure of WM's dynamic representations by maintaining individuation, and cannot be equated with the representational-units that carry bound information.

Michael Murez: *Thought coordination as concept identity*

William James (1890/2016, p. 459) observed that "the same matters can be thought of in successive portions of the mental stream, and some of these portions can know that they mean the same matters which the other portions meant." He claimed that this "sense of sameness is the very keel and backbone of our thinking" and argued convincingly that the "subjective sense of identity" between the objects of our thoughts is not explained by the objective identity between the "outer things", if any, those thoughts are about. James' "sense of sameness" anticipates the contemporary notion of coordination between thoughts (Fine 2008; Gray 2017; Goodman & Gray 2020, 2024; Murez 2023). "Coordination" (roughly) designates the relation between representations which the thinker is immediately disposed to treat as coreferential, without the mediation of any further representation of identity or coreference.

What, if not objective coreference, explains thoughts being coordinated or representing-as-the-same? A simple answer is that thought coordination is grounded in concept identity. According to this coordination-as-concept-identity view (henceforth, "CCI"), we represent things as the same by redeploying the same concept to think about them. Thus, we may conflate different things by representing them using the same concept (Landrum 2022), or conversely, represent one thing using different concepts, not realizing we are thinking of the same thing again (Frege 1892).

The main competitor of CCI is relationism (Fine 2008; Heck 2012, 2014; Goodman & Gray 2020, 2024; for review, see Gray 2017). Relationists deny that coordination is grounded in identity, equivalence, or any internal relation between representations, i.e., any relation determined by those representations' intrinsic properties. In particular, relationists deny that coordination between a pair of representations is explained by each independently having the same (or a similar) content or vehicle as the other. Coordination is claimed to be "irreducibly relational" and

grounded in informational processes which connect token representations of different semantic and vehicular types (Goodman & Gray 2020, 2024).

CCI can be fleshed out in different ways, by appealing to different notions of "concept". My goal will be to defend against relationism versions of CCI that construe concepts as robust representational vehicles. In doing so, I will draw on empirical considerations and focus on issues relevant not only to philosophers, but also to cognitive scientists.

First, I will clarify the explanandum. A theory-neutral way of understanding "coordination" is as shorthand for a complex of personal-level inferential/behavioral dispositions. I distinguish normative epistemological questions concerning the rational standing of those dispositions from descriptive psychological questions about the nature and (broadly) causal explanation of coordination. Among the latter, I further distinguish the question of how coordination dispositions are acquired diachronically from the question of which functional organization or mechanisms underlie them synchronically. I propose that CCI is best construed as sketching an answer to this last sort of question.

I then underscore a relationist insight: it cannot generally be assumed that the capacity to represent some relation R between the referents o1…on of some representational vehicles V1…Vn is grounded in the fact that R obtains between V1…Vn. This assumption is illegitimate even when R is identity: representing-as-the-same need not be grounded in sameness between representations. Yet the proper lesson to draw, I argue, is not that CCI is false, but only that its truth cannot be established a priori. The debate between CCI and relationism thus becomes an open empirical debate about how coordination actually works in human cognition.

We should seek accounts that secure "explanatory distance" (Taylor 2023) between their explanandum and the explanantia they offer. This speaks in favor of accounts which appeal to robust mental entities, i.e., entities (including processes/relations) whose characteristics and identity can, at least in principle, be ascertained through multiple means, rather than only based on whether thinkers superficially exhibit coordination dispositions. I argue that extant versions of relationism, and many versions of CCI, fail to satisfy this desideratum.

By contrast, versions of CCI that construe concepts as (re)identifiable by signature properties of their psychological and/or neurological vehicles make it possible to independently ascertain, through empirical means, when the same or different concepts are being deployed. I present two arguments – one theoretical, the other more empirical – in favor of such robust versions of CCI.

The theoretical argument amends influential regress arguments from Campbell (1988) and Millikan (1993). Supposing a plausible computational-representational theory of thought, there must be some relation between token thought-representations that does not itself need to be represented explicitly as the content of a dedicated representational vehicle. Instead, that relation must be encoded implicitly, in the sense that cognitive/computational processes are causally sensitive to its being instantiated by the vehicles themselves. Because identity between vehicles is guaranteed to be exploitable in the relevant manner by cognitive/computational processes, this makes it likely that such a relation should be exploited by well-designed cognitive systems to encode the ecologically crucial relation of identity between outer things.

The empirical argument develops the idea that relationism and CCI, if interpreted as making claims about robust mental entities, can be associated with competing empirical predictions. Unlike CCI, relationism predicts no syntactic (algorithmic level) or implementational (neural level) identity or resemblance between the vehicles of coordinated thoughts. For instance, relationism does not predict that personal-level coordination dispositions should reliably correlate with similarity in the patterns of neuronal activation that underlie separate deployments of concepts. I argue that such correlations in fact exist. As an illustration, I appeal to recent neuroscientific

research on the processing of pronominal anaphora, which is a paradigmatic way in which coordination between (singular) thoughts is linguistically expressed. Dijksterhuis et al. (2024) show that pronouns reactivate the same neuronal representations (known as "concept cells") as the antecedent nominals with which they are interpreted as coreferential. This is what CCI predicts, if concept cell reactivation is taken as a neurological signature of (singular) concept redeployment (Quiroga 2012). Beyond this particular empirical hypothesis, a more general lesson emerges: robust versions of CCI have powerful heuristic value as guiding hypotheses for multi-level empirical investigation of the "sense of sameness".

Tuesday 30th June 14:30 — Conceptual Development (Spiegelzaal)

Shruti Santosh: *Refining the role of analogy in Quinean Bootstrapping*

The rich array of usable concepts is a major factor in the power of adult human thought. But how do we come to learn concepts like INTEGER and CAUSE? Susan Carey's (2009) theory of Quinean bootstrapping is often invoked to explain the radical conceptual change in early childhood cognitive development that allows children to acquire such concepts. However, despite its influence, the mechanisms of Quinean bootstrapping remain somewhat obscure (Fodor, 2010). Furthermore, critics argue that the learning mechanisms cited by Carey, such as analogical reasoning, cannot perform real epistemic work in Quinean bootstrapping unless the relevant analogical mappings are already understood, leading to a vicious circularity (Reys, 2014). These critics argue that since the learner must already understand the target concepts and relations in order to map them, analogical reasoning cannot genuinely induce conceptual change. Defenders of Quinean bootstrapping sometimes respond by appeal to representational, computational, or architectural constraints that guide learning without the explicit representation of the target concept (Beck, 2017). This paper offers a different proposal: analogical reasoning contributes to Quinean bootstrapping primarily as a theory-generative mechanism that expands the learner's representational possibilities by generating new structural schemes. This contribution does not require "understanding" in the critics's strong sense (semantic interpretation, inferential mastery, or explanatory grasp).

I draw on a multi-functional view of analogy which is the idea that analogical reasoning serves different epistemic functions like theory generation, explanation, and confirmation, each with distinct representational demands and success conditions. I argue that if we focus on the role of analogy in theory generation, the circularity objection can be defused because analogies for theory generation don't require understanding in a strict sense. On this view, it is a mistake to impose the demands of explanatory or confirmatory analogies onto theory-generative analogies, which I will argue is the epistemic function analogy performs in concept development. I begin by describing the cognitive science of analogical reasoning by focusing on the dominant theoretical framework in cognitive science: Dedre Gentner's structure-mapping theory (SMT). SMT provides a precise characterization of what analogical reasoning is and how it could contribute to Quinean bootstrapping. Gentner (1983) proposes that analogy is fundamentally a mapping of relational structure from a base domain to a target domain. For instance, the analogy between the solar system and the atom maps the relational structure (the sun ATTRACTS the planets; the nucleus ATTRACTS the electrons) while discarding object-level properties (the sun is HOT; the nucleus is SMALL). The key insight is that analogies preserve relations between entities while discarding object-level properties. This is significant for the Quinean Bootstrapping thesis because it demonstrates that the selection of what gets mapped is governed by structural properties of the representation not by the reasoner's prior semantic understanding of the target domain.

Next, I apply this framework to Carey's flagship case of Quinean Bootstrapping, the acquisition of the concept of NATURAL NUMBER. As Carey presents it, Quinean Bootstrapping involves interactions between placeholder systems (e.g., the memorized count list), culturally transmitted

routines (counting practices), and learning mechanisms that enable the construction of new conceptual resources. I focus on fine-tuning the role of one learning mechanism already cited by Carey: analogical reasoning. I clarify what it means for analogy to "do work" in bootstrapping and reframes the circularity objection as a misallocation of epistemic burdens. I argue that the memorized count list is crucial because it can function as an externally scaffolded relational structure that helps satisfy the high demands of abstract analogical alignment. Even though the count list is not understood as representing numerosity, it provides a stable successor-like sequence that can be aligned with other structured domains (spatial paths; temporal sequences). These alignments support a theory-generative transition where the learner can now entertain structural hypotheses linking "next" to "one more," before they grasp the "cardinal principle" necessary for the concept of NATURAL NUMBER. This predicts familiar developmental dissociations such as rote counting prior to understanding and partial mastery of "next" preceding full numerical understanding.

I defend the plausibility of "analogy without understanding" by citing developmental evidence that young children can analogically reason about domains they do not yet fully comprehend. For example, Kotovsky and Gentner (1996) investigated children's ability to recognize relational commonalities. Crucially, they introduce progressive alignment as a learning mechanism the process by which attribute-based comparisons scaffold more abstract, distant ones. Children who first compare highly similar cases develop the representational resources to handle increasingly abstract analogies. This mechanism is relevant to Quinean bootstrapping as it shows how a child's representational repertoire can expand gradually through iterated comparison, without requiring the learner to possess abstract understanding from the outset. Likewise. Rattermann and Gentner (1998) demonstrated that relational labels improve young children's analogical performance. In cross-mapping tasks where object similarity conflicted with relational similarity, teaching children relational labels ("Daddy/Mommy/Baby" for monotonic size change) enabled them to preserve relational mappings despite misleading surface cues. Importantly, these labels do not need to convey deep conceptual understanding since they function as invitations to attend to relational structure. This connects to Quinean Bootstrapping's emphasis on how externally scaffolded structures like the count list can support analogical alignment before semantic understanding is in place.

Finally, I further motivate the idea that theory generative analogies do not require prior understanding by drawing a parallel between analogical reasoning to cases of conceptual change in the history of science, especially Nancy Nersessian's (2008) account of exploratory model-based reasoning. In scientific concept formation, analogies and models are often manipulated productively prior to settled semantic interpretation such as Maxwell's mechanical analogies in the development of field theory and Kepler's use of analogy in the development of his theory of planetary motion. The upshot is a more refined picture of analogical bootstrapping where theory-generative analogical reasoning introduces and recruits new structural organizations that only later promotes understanding. In this way, the paper defuses the circularity objection and sharpens the positive explanatory role of analogy in Quinean bootstrapping without inflating its role in concept acquisition.

Tobias Grossmann: *Childhood First: How Extended Development Made Us Human*

The evolution of human childhood presents a profound biological paradox. Human offspring are born in a state of secondary altriciality—helpless, immobile, and nutritionally dependent for years despite belonging to a lineage of otherwise precocial primates—yet this extended immaturity gives rise to cognitive capacities that are, in important respects, unique among animals: cumulative culture, language, and shared intentionality. For decades, the dominant paradigm in

evolutionary anthropology and psychology has held that this extended childhood is a metabolic consequence of encephalization. This "brain-first" paradigm posits that humans develop slowly because we are building a metabolically expensive organ; extended dependency is the price paid for intelligence.

This paper develops and presents converging fossil, comparative, and neurobiological evidence in support of an alternative account: the "Childhood-First Hypothesis." This hypothesis inverts the traditional causal narrative, proposing that extended childhood was not a consequence of brain expansion but its precondition, established in small-brained hominins through the social mechanism of cooperative breeding. The Decoupling of Brain Size and Developmental Timing The brain-first model relies on the Expensive Brain Framework, which predicts that developmental timing must scale with brain size due to metabolic constraints. Fossil evidence now decisively challenges this coupling. Synchrotron microtomography of the Dikika child (Australopithecus afarensis, ~3.3 Ma) reveals prolonged brain growth comparable to modern humans despite an ape-sized brain (~275 cc). Analyses of early Homo at Dmanisi (~1.77 Ma) demonstrate modern-like dental delays and a dentition growth spurt occurring at 5.3 years—intermediate between chimpanzees and humans—despite brain volumes one-third to one-half modern human size. Recent histological analysis of Homo naledi dental enamel reveals human-like growth rhythms in a species with australopith-sized brains (465–560 cc) that persisted until the late Middle Pleistocene. These findings indicate that the "small brain, slow life" configuration was a stable evolutionary strategy maintained for over three million years, not a transient stage. Developmental slowing preceded brain expansion.

Cooperative Breeding as the Evolutionary Driver

If metabolic demands of large brains did not drive this deceleration, what did? I propose that cooperative breeding acted as the evolutionary entry point into the human adaptive complex. Geochemical analysis of Australopithecus africanus teeth indicates that weaning occurred early (6–9 months), yet individuals survived severe seasonal stress—implying active allomaternal provisioning. Evidence from Dmanisi of an edentulous adult surviving for years without the ability to chew confirms that cooperative provisioning was operational in early Homo. Results from agent-based evolutionary simulations demonstrate that populations can evolve significantly longer childhoods without corresponding brain expansion when fitness benefits depend on alloparental investment. Cooperative care buffers the mortality costs of slow growth, creating a demographic niche where extended development becomes sustainable independent of encephalization. The "social cradle" of cooperative care transformed infant helplessness from a survival liability into an opportunity for prolonged social learning.

Neurobiological Implications: Extended Plasticity in the Social Brain

The Childhood-First Hypothesis generates specific neurodevelopmental predictions. If extended childhood preceded encephalization, we should expect prolonged plasticity in neural circuits supporting social cognition—and this plasticity should be linked to cooperative breeding rather than brain size. Evidence supports both predictions. In humans, the medial prefrontal cortex—critical for self-referential processing, mentalizing, and social decision-making—exhibits protracted structural and functional development extending into the third decade of life. Sensitive periods for social learning remain open far longer than in other great apes. Critically, comparative neuroimaging of common marmosets (Callithrix jacchus)—small-brained obligate cooperative breeders—reveals a strikingly convergent pattern: prolonged postnatal development of prefrontal and temporal-parietal regions relative to other primates of comparable brain size. This convergence suggests that cooperative breeding selects for extended plasticity in social brain

networks independent of overall encephalization—an adaptation for navigating the complex, variable relationships inherent to cooperative caregiving.

Cumulative Culture and the Ratchet Effect

This framework connects life history to cultural evolution. Extended childhood provides the necessary temporal scaffold for cumulative culture. The ratchet effect—where cultural innovations are preserved and elaborated across generations rather than lost—requires high-fidelity transmission. Extended childhood allows learners to thoroughly acquire complex skills and traditions before contributing innovations, preventing the cultural slippage observed in other species. The cognitive capacities required for cumulative culture—imitation, teaching, shared intentionality—are themselves developmental achievements enabled by the time and social input that cooperative breeding provides.

Conclusion: Toward an Interdisciplinary Research Program

I propose a two-stage model of human cognitive evolution. Stage 1 (Reorganization Without Expansion, ~3.3–1.8 Ma) involved the establishment of cooperative breeding and extended childhood, leading to neural reorganization—particularly in prefrontal regions—and the emergence of proto-cultural capacities in small-brained hominins. Stage 2 (Encephalization, ~1.8 Ma–present) saw rapid brain expansion driven by selection pressures generated by the increasingly complex cultural environments established in Stage 1.

This synthesis challenges accounts that treat human cognition as the straightforward product of a larger brain. Our distinctive cognitive profile is fundamentally relational constituted by networks of care that sustain prolonged development and the extended neural plasticity this affords. The Childhood-First Hypothesis raises questions that extend beyond any single discipline: questions about the nature of cognitive development, the relationship between biology and culture, and what it means for minds to be socially constituted. I offer this framework as an invitation for interdisciplinary engagement—from philosophers, psychologists, anthropologists, and neuroscientists alike—to test, refine, and extend these ideas.

Chiara Carraro and Manuel Bohn: *Mutual Exclusivity across Word Classes and Languages*

As children are exposed to a novel word it is often the case they must identify the correct referent among multiple potential ones. When a new word is uttered in the same context of a known and an unknown object, research provides unanimous evidence: children map the novel word onto the novel rather than the familiar object. This phenomenon is mostly referred to as Mutual Exclusivity (ME) and described as an important mechanism driving word learning (Markman & Wachtel, 1988).

However, there are different accounts as to what is the nature of the processes behind ME (Lewis et al., 2020): Lexical constrained accounts, which describe ME as a 1-to-1 lexical bias leading children to think an object can only have one label, are often tested against social pragmatic accounts, which attribute ME to inferential reasoning about the speaker's intention, logical inferences accounts, which attribute ME to logical reasoning, and probabilistic computational accounts, which attribute ME to probabilistic associative mechanisms. All these theories predict the same outcome, and therefore tend to be hardly falsifiable, despite the great variability in the paradigms employed in the literature. While extensively studied for nouns, little is known about novel verb disambiguation in an ME paradigm. Actions can, to some extent, be described with one basic-level label (the most efficient way to convey "eating" is <eat>, not <devour>) (Győri, 2019; Zhuang & Lingnau, 2022). This should lead children to reject second labels for already labelled actions. Research reports 2- to 4-year-olds being able to map novel verbs onto novel

actions successfully, but with accuracy decreasing with age (Merriman et al., 1993, 1996). The authors justify this finding referring to an increased experience with label overlap for actions (e.g., "running" can also be <jog>), which would lead them to accept second labels for familiar actions. No ME studies, moreover, were conducted on adjectives. The main function of adjectives is to discriminate between different versions of the same entity, making its use heavily dependent on the context (Tribushinina et al., 2013). In fact, an object can be described with multiple adjectives (a ball can be <round> and <blue> and <cold>), if no context is taken into consideration.

However, if one would refer to a specific object using an adjective, they would likely do so a) because alternatives are present in the context; b) referring to the feature of the target object that differentiates it from the non-target one. Inferences about the speaker's intention are therefore necessary to interpret adjectives, making their disambiguation more dependent on context and inferential reasoning about the speakers' potential alternative utterances and intention, rather than on 1-to-1 mapping between feature and label (a still necessary but not sufficient piece of information). Furthermore, no study has compared performance in nouns, adjectives and verbs in an ME paradigm, and none of the above-mentioned accounts makes differential predictions. Thus, this pre-registered study investigates how children's performance is modulated by word classes differently bound to basic-level labels and 1-to-1 mappings, and by the referent being an object, feature or action, with the goal to disentangle the influence of lexical and inferential processes in the classical ME paradigm. Additionally, we test the same predictions in a language with a less clear noun bias (Gözütok, 2024), Turkish.

Firstly, we predicted the ME-effect would be stronger for nouns than for adjectives and verbs: The noun condition represents a simpler condition, as nouns are more often referred to with basic-level labels, and words from basic-level categories are strongly mutually exclusive (Au & Glusman, 1990).

Secondly, we predicted adjectives would be easier than verbs for two reasons: Adjectives lend themselves, as per their function, to a more contrastive, contextual and inferential interpretation in comparison to verbs; Verbs describe actions, which are dynamic in nature, decomposable in other sub-actions, and therefore more complex. On each experimental trial, children see a character and three pictures, always consisting of two familiar objects (noun condition), features (adjective condition), or actions (verb condition), and one novel object, feature, or action (respectively). In experimental trials, the character asks the child to click on a picture using an unknown word, hence the target is always the novel picture; In filler trials, a familiar word is used to refer to one of the two familiar pictures. Each child completes 6 items per condition plus 12 filler trials. Results on 99 German children (mean age= 5.13; range: 3.05-6.85) show that children get better with age (β= 0.64, 95%CrI [0.23, 1.07]), a greater accuracy for nouns compared to adjectives (β= -1.18, [-2.22, -0.21]) and verbs (β= -2.31, [-3.34, -1.36]), and a greater accuracy for adjectives in comparison to verbs (β=-1.13, [-2.13, -0.13]); On an individual level, there were moderate correlations between adjective and noun (r= .34, p<.001), noun and verb (r= .32, p= .001) and a weak correlation between adjective and verb (r= .22, p= .032) accuracies.

Preliminary results on 56 Turkish children (preregistered sample= 100; mean age= 4.54; range: 3.20-6.29) also show better performance with age (β= 1.61, [0.91, 2-37]), a greater accuracy for nouns compared to adjectives (β= -1.24, [-2.36, -0.17]) and verbs (β= -2.35, [-3.48, -1.21]), and a greater accuracy for adjectives in comparison to verbs (β= -1.11, [-2.27, -0.03]); Performance above chance in all three conditions in German, the higher accuracy on adjectives in comparison to verbs and the age effect show that: Children succeeded in conditions when 1-to-1 mapping was not possible or not informative; They performed better in a condition where inferential reasoning was embedded in the word class function; Preschoolers' accuracy increased even with increased label overlap experience. In sum, our findings support the interpretation of ME as a flexible mechanism, strongly driven by inferential processes and modulated by the inherent lexical

characteristics of the word category. Final results on Turkish and theoretical implications for accounts of ME and word learning will be discussed.

Leonie Baumann, Lydia Paulin Schidelko, Marina Proft, Tanya Behne and Hannes Rakoczy:
Modal reasoning abilities emerge more unified than previously assumed

Living in an uncertain world, we represent and prepare for different possibilities all the time. It may or may not rain later, so should I bring an umbrella? We may or may not have milk left in the fridge, so should I buy some while in the supermarket? When confronted with such alternative possibilities, we can experience different kinds of uncertainty: with physical uncertainty the outcome has yet to happen, while with epistemic uncertainty the outcome has happened, but is unknown. From the perspective of cognitive development – mainly ontogenetically but in principle also evolutionarily one central question is: How do the capacities to represent these types of uncertainty develop?

One set of empirical findings suggests an interesting asymmetry in acquisition, which has, in turn, played an important theoretical role for accounts on the development of modal cognition. In the so-called “Doors task”, Robinson et al. (2006) asked 4- to 6-year-old children to plan in light of possible outcomes of events that were either about to happen or had already happened. Children had to place trays to catch a colored block that an experimenter pushed through a door of the corresponding color. In the physical uncertainty condition, children placed the tray(s) before the experimenter drew from a bag that contained orange and green blocks. In the epistemic uncertainty condition, children placed their tray(s), after the experimenter had already covertly drawn a block (Robinson et al., 2006).

To be sure to catch the block, children had to place two trays – irrespective of the type of uncertainty. However, while already 4- to 5-year-olds placed two trays in the physical uncertainty condition, even 6-year-olds struggled to do so in the epistemic uncertainty condition. From the point of view of various theoretical accounts on modal cognition, the purported developmental asymmetry between representing physical and epistemic uncertainty has substantial theoretical repercussions (e.g., Beck et al., 2012; Gautam et al., 2019; Phillips & Kratzer, 2024). For instance, Gautam et al. (2019) explain the asymmetry in acquisition by highlighting the role of the temporal structure of the possibilities. Their account proposes that while physical uncertainty involves representing how possibilities in the future may unfold from the present (from here onwards, the block that is drawn may be orange or green), epistemic uncertainty involves representing possible versions of the past (in the past, the experimenter could have drawn an orange or a green block…) and the resulting different versions of the present (... and as a consequence the block may now be behind the orange or the green door). They claim that this entails different levels of recursive meta-representation and thus different levels of complexity: representing one temporal junction for physical uncertainty (from the present to possible futures), but two temporal junctions for epistemic uncertainty (from the present to possible pasts, and from there to possible futures, of which one became the actual present) (Gautam et al., 2019).

These developments parallel the development of Theory of Mind – from first level to recursively embedded higher levels. Both representing physical uncertainty and representing first-order mental states (such as beliefs) are forms of meta-representation and thus co-emerge around age 4. One level up, representing epistemic uncertainty and representing second-order mental states (such as beliefs about beliefs) both are forms of second level recursive meta-representations and should thus develop in tandem at around age 6 (Gautam et al., 2019). Given the heavy theoretical lifting that is assigned to the assumed empirical asymmetry, it is surprising how little we know about its robustness and generality. In fact, on second sight, the current empirical situation turns out to be rather complex and puzzling. Recent findings from comparative and developmental

studies on modal cognition present a complicated picture with no apparent systematic pattern regarding epistemic and physical uncertainty. Against this background, the current studies aim to test for the robustness and generality of the asymmetry and its proposed relation to Theory of Mind.

In the project, we explored whether reasoning about physical and epistemic uncertainty follows distinct developmental trajectories by investigating children's ability to prepare for multiple incompatible outcomes under epistemic and physical uncertainty. In two preregistered studies, we explore whether young children treat physical and epistemic uncertainty differently when preparing for incompatible possibilities in closely matched minimal contrast pairs of uncertainty conditions using a within-subjects design (N = 202, 3- to 6-year-old German-speaking children, 95 female, 107 male). In the task, children were shown a set-up with two slides, balls and wagons. Children prepared to catch one ball whose trajectory was either known (control trials), unknowable (physical condition) or unknown (epistemic condition) by pushing one or two wagons under the slides. We further explored whether children's performance in the uncertainty task was related to their first- and second-order false belief understanding.

Across both studies, we did not find a developmental asymmetry: children did not prepare differently for physically and epistemically uncertain outcomes. Children seemed to either consider multiple incompatible possibilities in each situation or failed to do so altogether. Moreover, performance in the uncertainty task was not positively correlated with Theory of Mind. Taken together, these findings challenge two theoretical assumptions: that physical and epistemic uncertainty pose different representational demands and that they rely on different levels of recursive embedding that are related to children's developing Theory of Mind. Instead, our studies suggest that the ability to represent incompatible possibilities more unified than previously assumed and thus stand in conceptual tension with recent theoretical work on the development of modal cognition.

Tuesday 30th June 14:30 — Mental Health & Psychiatry (Voorkamer)

Benedetta Cogo: *Keeping it Rational: on Persons and Rationality in Philosophy of Psychiatry*

In this talk I emphasise the need, in philosophy of psychiatry, to bring some clarity on the role played by rationality in understanding mental disorders. This inquiry is framed within the idea that mental disorders are conditions that affect a person in their entirety, and that a proper understanding of the role of rationality is needed if theorising about psychiatry ultimately aims at furnishing insights into the factual, lived reality of psychiatric conditions. In fact, rationality has long been considered a central feature when characterising what constitute persons and the way they function, and a pivotal element when discussing mental disorders. I ultimately argue that rationality is an important feature that should be recognised and accounted for when theorising about mental disorders, although it is important not to overly emphasise its role as the identifying feature of the person. In particular, I focus on one significant tendency in the philosophy of psychiatry to adopt an approach that, going back to Davidson (1985, 2002) considers rationality as a constitutive feature of persons, without which we simply cannot make sense of our interlocutors as persons altogether.

I develop my case in four steps. First, I present George Graham's (2020) account of mental disorders as an exemplary approach that gives room to rationality without downplaying other important factors. I emphasise that Graham's framework is attractive insofar as it recognises rationality as a matter of degree rather than an all-or-nothing capacity, thereby offering a more nuanced understanding of mental disorders. I then proceed to show that Graham's framework draws on a Davidsonian view of rationality, according to which every intentional attitude must be to some degree rational. Second, I observe that the Davidsonian position according to which all

intentional attitudes must exhibit some degree of rationality for someone to qualify as a person seems to be, prima facie, too strong. I present three counterexamples in which intentional attitudes are not responsive to norms of rationality in different ways and yet we would not question, because of this lack of rationality, that they belong to a person. The cases I present are the following: delusions, belief-like states typically understood, at least to a certain extent, in terms of their irrationality; conspiracy theories, which constitute interesting cases precisely because of their irrational recalcitrance to counterevidence; and aliefs, arational, affect-laden intentional attitudes that happen as automatic responses and generate actions that are "belief-discordant" (Gendler 2008). Third, I consider a defence of a Davidsonian approach to delusions which shows in what sense a rationalist account of intentional agency is valid and in fact advantageous when accounting for delusions (Reimer 2011).

I go back to my examples and show that this counter-objection also applies to conspiracy theories. However, it does not seem to hold in the case of aliefs. Fourth, admitting that the notion of alief is controversial, I observe that even giving it a very minimal reading, the important difference with the other cases is that aliefs are not just irrational attitudes, they are "a-rational". In other words, they happen beyond the scope of the rational. The Davidsonian approach to rationality, being originally an interpretivist view, focuses on propositional attitudes, which are formulated within the domain of rationality (whether they are rational or irrational). However, I argue that there are reasons to distinguish intentional attitudes from propositional attitudes, where the former are not always bounded to norms of rationality. To show the importance of this distinction, I present the example of Mrs T, a woman affected by a neurodegenerative disease who progressively loses her capacity to formulate propositional attitudes (Stich 1996). Nonetheless, she is still capable of intentional interactions with the world. The example shows that all the intentional attitudes that are not propositional attitudes are not necessarily rational and that, following a Davidsonian rationalistic approach to intentional attitudes, we would need to not consider Mrs T a person.

Victor Lange and Sebastian Watzl: *Demons in depression: rumination and agency*

The literature of clinical psychology stresses that rumination, which involves repetitive thinking, is a central factor in major depressive disorder (MDD) (Watkins and Roberts, 2020; Wells, 2011). Recently, philosophers have begun discussing various issues concerning the agency of such rumination (e.g., Fanti Rovetta, 2025; Degerman and Sul, 2025; Farokhi, 2025). One central question seems to be: Are subjects active or passive in relation to their rumination in depression, that is, is rumination in depression something that subjects do or something that merely happens to them? Depressed subjects often describe their own rumination as involuntary and beyond control (Watts et al., 2017; Ciobotaru et al., 2024; Mancini et al., 2024), yet it remains an open question whether this description makes sense in the light of further philosophical examination. Answering this question is also of clinical importance. Current psychotherapeutic interventions train agential capacities such as attention control, introspective awareness, intentions, and values with the aim of strengthening a subject's ability to competently regulate their own rumination (Hayes, Strosahl, and Wilson, 2011; Wells, 2011). A successful philosophical account on the agency of rumination would likely inform such therapeutic protocols.

In a recent publication, Glasser and Irving (forthcoming) have defended a novel account on the agency of rumination – we can call it the 'affect in action account' (inspired by the title of their paper). This account makes two interrelated claims, namely: The rumination of depression involves the presence of so-called occurrent agency but the failure of so-called aggregative agency. Occurrent agency concerns an agent's capacity to guide mental and bodily processes as they unfold over time, while aggregative agency concerns an agent's capacity to organize and distribute actions over time. According to Glasser and Irving, this account explains why rumination

in depression might feel passive (by the failure of aggregative agency) while it is, in fact, active (by the presence of occurrent agency).

However, we shall argue that the affect in action account is problematic in various respects. Furthermore, we shall utilize the major points of our critique to develop the foundation for an alternative account. Concerning our critique, the affect in action account presents three motivating considerations for why the rumination of depression involves occurrent agency.

First, it claims that (1) subjects have personal-level attentional guidance of their rumination. To motivate (1), the account develops and draws on a thought example, Peter and his boss, as core evidence. However, as we shall argue, this example does not describe the rumination in depression adequately. The clinical literature stresses that the rumination in depression has the features of so-called brooding, involving negative valence, abstract construal, and cause-effect focus among other things. Instead, the example with Peter and his boos seems to describe another type of rumination, called pondering, which is not the type characterizing depression (Watkins and Roberts, 2020).

Second, the account claims that (2) subjects endorse their rumination through relevant goals. To motivate (2), the account refers to studies and models in psychology suggesting that rumination occurs as a response to goal-discrepancies, in particular higher-level goals (Martin and Tesser, 1989). However, even though the rumination of depression might be triggered or initiated by detected goal-discrepancies, the unfolding episodes of rumination are themselves not structured by these relevant goals. Rumination in depression does not exhibit the flexibility that commonly characterize goal-driven activity (Hayes, Strosahl, and Wilson, 2011).

Third, the account claims that (3) subjects can resist their rumination. To motivate (3), the affect in action account states that since subjects can resist other paradigmatically strong desires, such as those involved in addiction and Tourette's syndrome, and since such desires are stronger than an agent's desire to ruminate, then we have strong reasons to believe that depressed subjects can resist their rumination. However, as we argue, this transitive argument fails because there is an important difference in the relevant control structures.

Utilizing the major points and insights of our critique of (1)-(3), we then lay the foundation for an alternative account of the agency of rumination in depression. We call this the 'managerial control account' (inspired by the terminology of Hieronomy's, 2006, discussion of doxastic control). This account stresses that episodes of rumination in depression are not themselves guided or controlled 'from within', but subjects can guide and control their responses to the occurrences of these episodes, hereby indirectly influencing the causal impact of the rumination and future occurrences. We draw on the work of third-wave cognitive-behavioural therapy, which strongly seems to stress that subjects cannot directly control their ruminative thoughts, though they can engage in certain synchronic and diachronic control structures to influence the cognitive and emotional effects of these thoughts (Hayes, Strosahl, and Wilson, 2011; Wells, 2011).

To develop this account further, we introduce two other considerations from the empirical and clinical literature. First, we stress that rumination in depression is the result of an extended habituation process. However, the rumination of depression is a particularly strong habit in the sense that it is unusually difficult to change (Watkins and Roberts, 2020). Second, we stress that rumination in depression does exhibit the features of a kind of self-narrative (Fanti Rovetta, 2025), though compared to self-narratives commonly, rumination is more rigid in its content.

We close our presentation by some concluding remarks on the observation that depressed subjects often describe their depressive thoughts and feelings as demonic (e.g., Solomon, 2014). Drawing on inspiration from Søren Kierkegaard's notion of 'the demonic' as enclosing and monological patterns of thinking and feeling, as in The Concept of Anxiety (1844 [1980]), we

stress that the 'managerial control account' offers an interesting perspective to make sense of how 'demons in depression' develop and unfold in the form of rumination.

Mara McGuire and Joshua Kramer: *Does everyone know what addiction is?*

The philosophy and psychology of addiction are at an impasse. For decades, researchers have focused on solving the so-called "puzzle" of addiction: why do addicts use drugs despite negative consequences (Pickard 2021)? Two types of solutions are common, one appealing to compulsion (Leshner 1997; Charland 2002; Holton and Berridge 2013) and another to choice (Pickard 2012; 2015; 2020). Neither solution has proven satisfactory and, recently, theorists have started to incorporate further constructs into their solutions, such as community- and self-identity (Sinnott-Armstrong and Pickard 2013; Pickard 2012, 2015, 2020). But even if we find a satisfactory solution, the puzzle ignores many important and overlooked features of addiction: (i) relapse after extended periods of abstinence, (ii) resumed drug use on "special occasions," (iii) the role of curiosity as a risk factor for addiction, and (iv) the high comorbidity of addiction and attention disorders (like ADHD) (Zuckerman 1994; Kashdan et. al. 2004; Lindgren et al. 2010; Anderson et. al. 2011; Leslie 2014; Heyman 2019). These dimensions of addiction not only elude compulsion- or choice-based explanations, but also suggest a central role for attention in understanding addiction. Thus, we ask: might addiction be a problem of attention?

We are not the first to detect a crucial link between addiction and attention. In The Principles of Psychology, William James (1890) considers whether attention may be central for understanding addiction: "being a drunkard…that is the conception that will not stay before the poor soul's attention" (565). While James never pursued a full-blown account of addiction—let alone one centering addiction—passing comments like this one, combined with his claim that "what holds attention determines action" (1892: 319), strongly suggest that James views addiction as primarily a problem of attention, rather than of conscious choice, motivational ideals, or compulsion. In fact, his idea has left a very tangible effect on contemporary addiction treatment. Alcoholics Anonymous (AA), the ubiquitous addiction treatment plan developed in the 1930s, credits James as a primary influence on their conception of addiction (Bevacqua and Hoffman 2010). In scientific and philosophical contexts in subsequent decades, however, James's tantalizing suggestion that addiction is fundamentally a problem of attention has been curiously absent.

No one denies that attentional processes are part of addiction's story; passing comments to the effect that attention is somehow involved in addiction are frequent in the literature (Henden 2019; Kennett et al. 2019). But we posit that addiction is a problem of attention in a stronger, more fundamental sense. We first show that compulsion- and choice-based explanations implicitly depend on more fundamental commitments regarding attention. For example, advocates of the incentive-salience account, a compulsion-centered view, refer to external cues as "capturing" the attention of the addict (Berridge and Robinson 2017; Anderson, Laurent, Yantis 2011; Field and Cox 2008). On the other hand, advocates of the choice-based explanation focus on addicts' attending to their person-level values or self-conceptions in a reflective, deliberative way (Pickard 2012, 2015, 2020). These types of attention, which we call environmental and reflective attention, respectively, are certainly part of the story. But they cannot explain the important, yet overlooked, phenomena mentioned above (e.g., "special occasions" of relapse, high curiosity as a risk factor for addiction, co-morbidity of addiction and ADHD).

We then develop a novel account of an additional mode of attention: the organizational mode. Drawing on work on perspectives (Camp 2019) and an analogy with moods, we characterize this organizational mode as a disposition toward different ways of structuring how one is attending rather than what one is attending to. Unlike the environmental and reflective modes, the organizational mode is characteristically holistic and non-intentional; thus, organizational attention

can be understood as a style of attention in the way that “being joyful” or “being melancholic” are holistic frames or styles in the affective realm. We identify rigidity and flexibility as key dimensions along which styles of organizational attending can vary. A flexible organizational mode can structure, frame, and orient attending with an agility that is sensitive to the variably textured internal and external environment, as well as the agent’s shifting practical and theoretical interests. A rigid organizational mode of attending, on the other hand, assumes the same structuring, framing, and orientation of attention, regardless of changes in the internal and external environment or the agent’s interests. As with moods, this rigidity or flexibility is not voluntary and is, at best, distally controlled.

Ultimately, we argue that this organizational mode of attention is central for understanding addiction. Addiction is often described as a problem of rigidity over and above its given object. That is, addicts often experience a general feeling of being “enslaved” by a system of routines and habits that holistically anchor their life and that cannot be sufficiently explained by a more restricted, object-focused compulsion. Our picture explains this rigidity: an addict often too narrowly structures, frames, and orients attention in a way that non-addicts would acknowledge swings free from their richly and variably textured internal and environmental interests—e.g., lacking more beneficial stress coping-mechanisms, reframing tools, or modes of problem solving given goals. In addition, the organizational mode can better explain why the addict appears to lack voluntary control—a central desideratum of current theories of addiction. A non-intentional mode of structuring and framing attention is not easily accessible from the reflective mode of attention and is not related to the number of objects attended to (environmental mode of attention).

Our notion of a rigid organizational mode also offers novel insights into overlooked features of addiction. For example, empirical evidence suggests that “absorption,” an element of curiosity characterized by a tendency toward intense engagement, is a risk factor for substance abuse (Lindgren 2010; Kashdan 2004). Our account can uniquely explain this finding by pointing to absorption-curiosity as a manifestation of the rigid organizational mode of attention in addiction. Finally, we draw further connections between rigidity and the attentional difficulties characteristic of ADHD, laying the foundation for an account of addiction that can make sense of the still unexplained comorbidity of ADHD and addiction.

Anssi Bwalya and Polaris Koi: *Experiencing Agency in the Context of ADHD: The Role of Agential Capacities and Sociodemographic Factors*

Agency and related constructs, such as action control, self-efficacy, and autonomy, are widely studied topics in psychological and philosophical research (see Bandura, 2018; Gallagher, 2012; Haggard, 2017). Existing literature points to the importance of experienced agency in health and wellbeing, economic development, and political participation (e.g., Martikainen, 2025; Moore, 2016; Wuepper & Lybbert, 2017). However, the exact definitions of agency have varied across studies and subfields. Different theoretical and methodological approaches highlight different factors underlying agency – from neurocognitive mechanisms to societal structures. We propose an integrative perspective, according to which the subjective experience of agency is shaped by one’s agential capacities, such as self-control, decision-making, and motivation, as well as by one’s position in the society. That is, people’s experience of agency, or lack thereof, builds on their individual capacities and their subjective perceptions of those capacities, but this happens in dynamic interaction with their immediate environment and broader structural conditions. In the present study, we will follow this theoretical framework and collect empirical data on people’s subjective experiences of agency.

We will focus on two complementary aspects of agency: people’s sense of agency and their general self-efficacy. Sense of agency refers to people's perceived ownership of their actions,

"the registration that I am the initiator of my actions" (Synofzik et al., 2013; see also Tapal et al., 2017). General self-efficacy reflects a more goal-oriented facet of agency, people's beliefs about their ability to sort out challenges and reach their aims (see Bandura, 1997; Luszczynska et al., 2005). We will combine questionnaire measures and executive functioning (EF) tasks (i.e., cued task switching, n-back, and Stroop) to examine agential capacities and the subjective experience of agency in the context of Attention Deficit/Hyperactivity Disorder (ADHD). ADHD is characterised by impairments in EF, leading to attention difficulties, hyperactivity, and impulsivity. These traits create additional challenges for agential capacities. For example, intrapsychic self-control strategies that rely on effortful attention control may be less effective for people with strong ADHD traits (Koi, 2021).

Moreover, if ADHD traits lead to recurring challenges in daily life, these negative experiences may shape people's agency beliefs beyond the direct effect of their objective EF impairments. Indeed, previous studies offer preliminary evidence for a negative association between ADHD and general self-efficacy (Newark et al., 2016; Waite et al., 2022). That said, to our knowledge, no studies so far have examined ADHD-related differences in people's sense of agency rather than their self-efficacy. In addition to these cognitive and behavioural phenomena, we expect that people's experiences of agency are also shaped by their social position in terms of both socioeconomic status (SES) and gender. Many studies report positive associations between SES and domain-specific aspects agency, such as self-efficacy in science (Tan et al., 2023), career goal setting and exploration (Wang et al., 2025), and sexual life (Cha, 2022).

Furthermore, some research suggests potential gender differences in sense of agency (Hurault et al., 2020) and self-efficacy (Bonsaksen et al., 2018; Löve et al., 2012), such that men report higher agency than women. Finally, it is important to note the association between depressive symptoms on experiences of agency. Research literature on self-efficacy (see Li et al., 2024) and sense of agency (Borrelli et al., 2026; Di Plinio et al., 2024; but see also Bart et al., 2023; Tapal et al., 2017) suggests that depressive symptoms may undermine both aspects of agency – or vice versa. However, inconsistent findings regarding the association between depression and sense of agency highlight the need for further research.

In this study, we will examine the extent to which people's sense of agency and general self-efficacy are associated with their EF capacity, self-reported ADHD traits, depressive symptoms, SES, and gender. We expect that EF capacity, weaker ADHD traits, less depressive symptoms, higher SES, and male gender will each independently predict stronger sense of agency and higher general self-efficacy – even after controlling for the effects of the other predictors. We will test this hypothesis with multiple linear regression models. Additionally, we will carry out network analyses to explore the links between sense of agency, self-efficacy, specific EF tasks, ADHD traits, depressive symptoms, and demographic variables. The study plan will be preregistered in February, all data will be collected between March and April, and we will carry out our main analyses in May. The results of our study will offer new insights into how different psychological and sociodemographic factors shape the subjective experience of agency. In addition to more nuanced theoretical understanding, this work may provide valuable information for interventions aimed at supporting people's agency in disadvantaged circumstances.

Tuesday 30th June 14:30 — Logic & Rationality (Bovenkamer)

Konrad Rudnicki, Piotr Łukowski, Bert Leuridan, Vanessa Doreen Ruiz Stovel and Karolien Poels: *How do humans process contradictory cues - an EEG-ERP study testing hypotheses derived from classical and non-classical logics*

Research on the cognitive processing of contradictions has predominantly focused on rhetorical or context-dependent inconsistencies rather than strict logical contradictions of the form (p and

not-p). Consequently, empirical evidence regarding how the human cognitive system handles genuinely inconsistent information remains limited. This study addresses this gap by investigating the neural processing of true logical contradictions within a predictive coding framework, where logical conclusions are operationalized as sensory predictions. We derived competing hypotheses from classical and non-classical logic: (H1) the principle of explosion from classical logic, which posits that from a contradiction anything follows, implying the brain generates multiple, mutually exclusive predictions when facing contradictory premises; and (H2) the principle of implosion from the non-Fregean logic of content, which suggests that contradictions lead to a suspension of inference and an inability to form any clear predictions.

We conducted an EEG experiment with N = 30 participants using a visual oddball task. Participants learned associations between letter cues and target shapes (e.g., symbol "A" predicted a smooth shape, while "B" predicted a spiky shape). The experiment included consistent cues ("AA", "BB"), contradictory cues ("AB", "BA"), and novel/meaningless cues ("C"). Brain activity was recorded via 32-channel EEG, focusing on event-related potentials (ERPs) indicative of prediction errors and cognitive load: the visual mismatch negativity (vMMN), P3b, and the Late Positive Component (LPC).

The results showed that targets following consistent cues but violating learned associations elicited a significant vMMN at right-frontal electrodes (F8, FT10), reflecting a standard prediction error. Crucially, targets following contradictory cues did not elicit a vMMN, supporting the principle of implosion (H2) where the predictive system fails to generate specific expectations. Interestingly, novel/meaningless cues did elicit a significant vMMN, suggesting they were treated as cues for potential future events rather than total inferential dead-ends. Furthermore, both contradictory and meaningless cues elicited significantly enhanced P3b and LPC amplitudes compared to consistent trials, indicating that while contradictions may not generate predictions, they prompt the allocation of additional conscious resources for evaluation and working memory processing.

Exploratory analysis of individual differences revealed that a subset of participants (n = 13) employed explicit strategies to sidestep contradictions, such as focusing on only one part of the "AB" cue. For these individuals, neural activity revealed a significant vMMN when the subsequent target violated their self-reported strategy. This indicates that these participants effectively "removed" the contradiction to restore consistency and form single-cue predictions. However, even for strategy-congruent targets, a weakened prediction error was observed, suggesting that the effort to ignore contradictory information was only partially successful.

These findings suggest that for intramodal visual information, the human cognitive system tends toward implosion rather than explosion. While cross-modal studies suggest that conflicting predictions can coexist across different sensory streams, our results indicate that contradictions within a single modality disrupt the formation of predictive models. This supports the utility of non-classical frameworks, such as the logic of content, for modeling human information processing under inconsistency.

Beyond the specific empirical findings regarding contradictions, this study serves as a proof-of-concept for the broader feasibility of "psychologistic logic" - a framework that utilizes formal logical systems as descriptive tools to model the architecture of human information processing. By adopting a broad definition of reasoning that encompasses both deduction and unconscious sensory processing, the research demonstrates how logic can bridge the historical rift between formal philosophy and cognitive science. Central to this approach is the methodological use of paradoxes and contradictions as "edge cases". Much like extreme phenomena in natural sciences, logical anomalies in psychology can allow researchers to identify which specific logical

properties - such as paraconsistency or the principle of explosion - accurately reflect the constraints of human cognitive architecture.

Hanna Schleihauf, Emily Sanford, Bill Thompson, Josep Call, Esther Herrmann and Jan Engelmann: *Rational Belief Revision in Chimpanzees*

Rationality has long served as a central standard for evaluating human thought and has traditionally been regarded as uniquely human. Reasoning is typically considered rational only insofar as beliefs are grounded in, and appropriately updated by, the evidence available to the thinker. One particularly informative way to study rationality is by examining selective belief revision—that is, whether and how agents revise their beliefs as a function of the strength and relevance of new evidence.

In this talk, I present four studies suggesting that the capacity for rational belief revision is shared with our closest evolutionary relatives, chimpanzees. Across these studies, we investigate whether chimpanzees revise their beliefs in ways that are sensitive to evidence strength, second-order evidence, and social sources of information.

In the first two experiments, we investigated chimpanzees' (N = 15; preregistered) responses to counter-evidence. A piece of food was hidden in one of two locations. Chimpanzees first received evidence favoring one location and made an initial choice. They were then presented with a second piece of evidence favoring the alternative location and made a second choice. Crucially, we varied whether the second evidence was weaker (strong-evidence-first condition) or stronger (weak-evidence-first condition). Across both experiments, chimpanzees were significantly more likely to revise their beliefs when the initial evidence was weak than when it was strong (Experiment 1: $\chi^2 = 14.03$, df = 1, $p < .001$; Experiment 2: $\chi^2 = 20.52$, df = 1, $p < .001$), indicating sensitivity to the relative strength of competing pieces of evidence.

In the third experiment (N = 22 chimpanzees; preregistered), we examined whether chimpanzees revise their beliefs in light of second-order evidence, specifically undermining defeaters—information that weakens the evidential basis of a belief without directly supporting an alternative. The experiment included defeater and non-defeater conditions and was conducted in both visual and auditory modalities. In the visual defeater condition, after chimpanzees had chosen a location supported by visual evidence (seeing food through blurry glass), the experimenter revealed a screen with a picture of an apple glued to it, thereby undermining the original evidence. In the corresponding non-defeater condition, the revealed screen was fully transparent, leaving the original evidence intact. In the auditory version, chimpanzees first chose based on a rattling sound; the evidence was undermined by revealing a stone (defeater) or left intact by revealing a leaf (non-defeater). As predicted, chimpanzees were significantly more likely to switch to the alternative location in defeater than in non-defeater conditions ($\chi^2 = 16.27$, df = 1, $p < .001$), with this effect holding across modalities and no interaction between condition and modality ($\chi^2 = 0.11$, df = 1, $p = .743$).

In the fourth experiment, we tested whether chimpanzees (N = 22 chimpanzees; preregistered) revise their beliefs rationally in social contexts, when their own belief conflicts with that of a partner. Subjects first made a choice based either on evidence or no evidence. A competitor then made the opposite choice, also either with or without evidence. Importantly, subjects could observe whether the competitor's choice was evidence-based, even if they could not see the evidence directly. The results showed that chimpanzees integrated social and epistemic information in a rational manner: they revised their believes more often when they themselves

lacked evidence and the competitor had strong evidence than when the reverse was true ($z = 2.02$, $p = .043$).

Taken together, these findings demonstrate that chimpanzees selectively revise their beliefs in ways that are sensitive to evidence strength, the integrity of evidential support, and social sources of information. Rather than revising beliefs indiscriminately, chimpanzees adjust their beliefs in a manner consistent with core principles of rational reasoning. These results challenge the view that rational belief revision is uniquely human and suggest deeper evolutionary roots of this capacity than previously assumed.

Chenwei Nie: *Rational People's Irrational Beliefs*

In the middle of this brutal summer, with scorching heatwaves across Europe, Asia, and North America, it is poignant that so many believe that climate change is not real. Nearly 15% of Americans believe it is not real (Gounaridis & Newell, 2024), and among the members of the 118th US Congress, almost a quarter share this belief, who are all Republicans (So, 2024). Even more alarming, climate denial is just one example in a much broader, unsettling set of irrational beliefs, including but not limited to cases of superstitious, religious, political, and conspiratorial beliefs. The burning question is: Why do people obstinately hold onto irrational beliefs in the face of counterevidence?

When a belief fails to be sensitive to evidence as it should and is epistemically irrational (Bortolotti, 2010), it is likely influenced by non-evidential factors that do not bear on the truth of the belief. In the case of climate denial, the non-evidential factors may include the denier's anxiety over the catastrophic implications of climate change, her fondness for Trump whom she may take to be a climate denier as well, and her practical consideration that the belief solidifies her identity as a proud Republican. What is less clear is how exactly non-evidential factors contribute to the development of irrational beliefs.

Existing approaches often assume that non-evidential factors either distort the way the person collects and evaluates independent evidence or are taken by the person as evidence for her belief (Schleifer McCormick, 2020; Glüer & Wikforss, 2022; Simion, 2024; Flores, 2025).

While acknowledging that these irrational cases can occur, in this paper I will propose a new approach that does not presuppose such irrationalities and allows for the possibility that the person is clear-eyed about the conflict between her belief and independent evidence. According to this new approach, non-evidential factors may contribute to the formation and maintenance of a distinctive kind of seeming experience (also referred to as appearance, intuition, or impression). The climate denier's anxiety over the catastrophic implications of climate change, her fondness for Trump, and her practical consideration may give rise to a very strong seeming that climate change is not real. A seeming itself is not a belief (Pryor, 2000; Huemer, 2001; Moretti, 2020; McAllister, 2023). But a strong seeming can, in various ways and degrees, compel belief. Specifically, a strong seeming that p may causally incline the person to believe that p (Nie, 2025), may make the person think or feel that it is justified to believe that p, and may make the person think or feel that she simply knows that p.

As a result, the person may find herself in an epistemically dilemmic situation: on the one hand, she faces and, to varying extents, recognises the independent evidence against her belief, but on the other hand, she experiences a strong seeming that compels the belief. No epistemic norm is easily available for one to resolve this tension (for an inverse case where the seeming is true and the independent evidence is false, see Williamson, 2025). What the person ultimately believes depends on a complex interaction between her seeming and her consideration of the independent

evidence. The interaction may differ from case to case. Whenever a false seeming prevails, the person will end up with an irrational belief in the face of counterevidence.

Chiara Saponaro, Mahham Fayyaz, Grace Pavalko and Nicolò Cesana-Arlotti: *Logical thought and logical words: Developmental links between non-verbal reasoning and language*

Introduction: How are logical reasoning and language related in development? A core question concerns whether logical reasoning depends on language, or whether it can emerge independently and only later become integrated with linguistic representations[1-3]. Disjunctive reasoning, formalized as A ∨ B; exclude A → B, is a logical ability thoroughly studied by developmental psychologists. Recent work shows that preverbal infants can draw disjunctive inferences from visual events as evidenced by their sensitivity to violations of such inferences[4,5]. By contrast, from a linguistic perspective, preschoolers do not consistently assign adult-like interpretations to sentences containing the logical connective "or" until later in development[6,7], and its production is rare compared to other connectives such as "and"[8,9]. This raises the question of when children understand the meaning of "or", and how non-verbal abilities relate to the later acquisition of logical language. Specifically, can 3-year-old children use the connective "or" to derive disjunctive inferences? Moreover, can non-verbal disjunctive inferences scaffold children's comprehension of this logical connective?

Objectives and methods: To address these questions, we conducted two preregistered experiments with English-speaking 3-year-old children. Our aims were (i) to test whether children at this age can perform disjunctive inferences in non-verbal and linguistic contexts, and (ii) to examine whether prior engagement in non-verbal disjunctive reasoning facilitates children's interpretation of "or". In Experiment 1, 24 children (M age = 39 months 10 days; range = 36m12d – 42m22d; 14 boys) first completed a non-verbal inferential cognitive task involving only visual information. Two objects sharing an identical top were hidden behind an occluder; a cup scooped one object, revealing only its top; the other object then briefly reappeared from behind the occluder; finally, children were asked which object was in the cup. Success required excluding the visible alternative and inferring the identity of the hidden object (Figure 1A). Children then completed a linguistic inferential task, which was structurally identical to the first task but with one key difference: children heard a disjunctive statement (e.g., "The object is the car or the ball") and had to integrate it with visual evidence to identify the correct object (Figure 1C). In Experiment 2, a new group of 24 children (M age = 38 months 19 days; range = 36m9d – 42m17d; 11 boys) completed the same linguistic inferential task, but crucially, instead of first completing the inferential cognitive task, they were initially tested on a non-inferential control task in which the cup's content was directly revealed, thus not requiring disjunctive inference (Figure 1B). Each task consisted of four trials, and the experiments followed a fully counterbalanced design yielding 16 conditions crossing four factors: target object, order of linguistic disjuncts in the inferential linguistic task, object position on the screen, and trial order.

Results: All analyses were conducted using Bayesian one-sample Wilcoxon signed-rank tests (see Figure 1D). Children succeeded in the non-verbal inferential task in Experiment 1, identifying the correct object well above chance (M = 88.19%, SD = 24.32; BF10 = 6715). They also succeeded in the non-inferential control task in Experiment 2 (M = 95.49%, SD = 10.42; BF10 > 10,000). In the linguistic task, we distinguished two components of disjunctive reasoning. First, we tested whether children interpreted "or" as restricting the set of possible referents to the verbal disjuncts (i.e., from three initial objects to "A or B"). Children in Experiment 1 named one of the disjuncts significantly above chance (M = 88.54%, SD = 22.1; BF10 = 45.14), whereas children in Experiment 2 did not (M = 74.65%, SD = 23.38; BF10 = 0.542). Second, we tested whether children could exclude one of the two disjuncts based on visual information. Children in both

experiments succeeded on this component, reliably selecting the correct alternative once the relevant set was established (Exp1: M = 90.28%, SD = 16.6; BF10 > 10,000; Exp2: M = 89.24%, SD = 18.3; BF10 > 10,000). A direct comparison confirmed that children in Experiment 1 showed a stronger interpretation of “or” as an exhaustivity operator – that is, as a linguistic cue to define a set of disjunctive alternatives – than children in Experiment 2 (BF10 = 2011.87, ω = 0.75, 95% HDI [0.617, 0.876]; Bayesian Mann-Whitney test).

Discussion: Our results have major implications for the development of logical reasoning, the acquisition of verbal disjunction, and the relation between the two. First, we found that at age 3 children are not only highly proficient in deductive reasoning based on visual information but can also generate disjunctive inferences by integrating their comprehension of “or”-phrases with visual information. This suggests that the ability to draw deductive inferences may contribute to the early meaning of “or”[10]. Crucially, we also found that, at the onset of verbal disjunction, even a brief practice with the non-verbal disjunctive inference documented in infants[4,5] improves children’s comprehension of verbal disjunction. These findings provide strong support for the proposal that preverbal disjunctive inference functions as a developmental precursor scaffolding the acquisition of the meaning of “or”.

Tuesday 30th June 17:00 — 4E Cognition (Kerkzaal)

Elias Cohen: *Situatedness and intentionality: a salience-based approach to non-representational cognition*

Abstract: It’s 6am when the sound of rain hitting the window of your room wakes you up. The first thought that comes to you is: it’s raining again. Though you are in London, you don’t think that it’s raining in London again. Somehow, the mere fact of your presence in London suffices to make the truth of your thought, its content, rest on the location, without it being represented. This is the starting point of Perry’s (1986) famous analysis of unarticulated constituents: features of the truth-conditional content that are not mentally represented. What is more directly interesting for philosophy of mind is the distinction that Perry makes between two forms of intentionality: ‘aboutness’, which involves representation, and ‘concerning’, which consists in the sheer act of making the truth of the thought turn on a certain unrepresented, possibly unconceptualized, aspect of the environment.

The question then arises: what must the environment be like for it to be possible for the thinker to make the truth of her thought turn on some of its features in such a way? I would like to suggest that what makes it possible for you to hinge the truth of your thought on features of the environment that you do not represent, is determined by situation-specific salience lists. Indeed, I follow Barwise & Perry (1983) and Barwise (1989) in thinking that the thinker is never related to the environment itself, but to a part of that environment, which is determined by what her attention is restricted to (what Barwise calls ‘focus situations’). To be in a situation, then, is to activate certain ‘mental constraints’ (Breheny 2002) that make only certain thoughts and representations available. I propose to say that the situation sets appropriateness conditions for thoughts, which can then be assessed for truth or falsity. This means that appropriateness conditions act as the necessary frame for semantic evaluation. I would also like to propose some clarifications on the notion of salience.

I will distinguish between two varieties: cognitive salience and ontological salience (cf. Schmid 2007). Cognitive salience is relative to what has been actively raised to salience (what you endogenously focus your attention on), whereas ontological salience refers to what is salient for an organism by default (or what your attention is exogenously drawn to). For instance, it would appear that changes in animals are easier to detect than changes in inanimate objects in otherwise identical pictures of natural scenes (New, Cosmides & Tooby 2007). The proposal is

that precisely how salient the feature of the environment is, determines the extent to which it is possible to hinge the truth of thoughts and, perhaps, the success of actions, on it, without representing it. I will discuss some objections to this view and some applications as well.

Marco Facchin: *Comparative cognition is modestly, but radically, embodied*

Comparative approaches to cognition - the study of cognitive capacities across (animal) species - often presents itself in rather cognitivistic terms. Indeed - barring the occasional exception [e.g. 1] - these approaches may appear as the last refuge of classical cognitivism. For example, they still often conceives of representations as explicit, declarative and "action neutral" informational structures [e.g. 2], and still often takes cognition and association to be opposite and exclusive [e.g. 3] - in open contrast with much of "human", non-comparative cognitive science [e.g. 4].

My talk aims to dispel this impression, showing that some of the explanatory practices of comparative approaches to cognition are more naturally and usefully described as appealing to conceptual and experimental motifs [cf. 5] belonging to radically embodied (non-representational and non-computational) cognitive science. I will support my claim by analyzing a number of case studies. The first batch of case studies is taken from a recent book-length review of avian cognitive neuroscience [6]. I illustrate, using an appropriate number of textual quotes, how the review often claims that standard cognitivist posits (such as various types of "cognitive maps") are central to explanation of the behavior of birds - only to then illustrate how, in the actual explanations the review itself offers, cognitive maps are not at all mentioned, and all the hard explanatory work is carried out by complex, structured, multimodal and informationally rich environmental information.

This is in line with the conceptual motifs and explanatory preferences of radical embodiment, which characteristically appeals to the richness of environmental information to dispense from positing inner representations [5]. I will then examine some famous, and relevant, experiments in comparative cognition [7-10], and argue that they make use of experimental techniques and constructs that are central to radically embodied cognitive science, such as the identification of body-scaled, "dimensionless" pi-numbers to identify affordances [11,12] and the identification of patterns of sensorimotor contingencies and patterns of sensorimotor engagement to simply and resolve complex cognitive tasks [13,14].

The above cases show that radical embodiment has a role in comparative cognition. How big is this role? I'll suggest that it is modest, but non-negligible. It is non-negligible, in that it prompts us to be wary of the "classical cognitivist" rhetoric comparative approaches to cognition adopt, and revise our understanding of comparative cognition accordingly. But it is modest, in that it does not support any blanket anti-representational and/or anti-computational claim. With a slogan: comparative cognition is modestly, but radically, embodied.

Miguel Gramage: *Usage, Coordination, and the Modality of Ecological Information*

Ecological psychology (EP) explains perceptual behaviour in terms of affordances—the opportunities for action offered by the environment—and ecological information—the sets of structures in the ambient array that allow the animal to engage with those opportunities. Since information is taken to specify affordances, EP claims that animals can directly perceive opportunities for action by picking up patterns in the optic array (Gibson, 2015; Michaels & Carello, 1981). There is, however, a live dispute over how specificity should be understood. One influential proposal is the usage-based account. Unlike the orthodox view (Turvey et al., 1981), it rejects the idea that specification consists in a nomological relation between ambient patterns and the perceived properties of the environment. On the usage-based view, specification is instead a relational achievement between ambient patterns and the animal's activities. What matters, in

short, is how patterns of light, sound, and so on are used in goal-directed activity. Objects of perceptions are specified in use. Despite having recently attracted more attention, this picture has raised a familiar objection (Segundo-Ortin et al., 2019; Carvalho & Rolla, 2020). The worry concerns whether the usage-based account can accommodate a truly modal conception of ecological information. The objection insists that the informational significance of an ambient pattern derives from what it affords—the possibilities for action it offers—rather than from what some particular agent happens to do with it at a particular moment. Consider, for instance, a torch whose vibration intensity varies with proximity to an object. The vibration pattern plausibly remains informative regardless of whether the agent is currently exploiting it. As this suggests, the property of being informative is not exhausted by an organism's actual, occurrent uses. However, by insisting that information derives from the actual use of an ambient pattern, the usage-based account seems unable to account for the very idea of an ambient pattern being usable or affordable, and not only for its being used and afforded. Therefore, it seems we must appeal to something more than the mere use—such as a nomological relation of covariance—to explain the modal character of ecological information.

In this paper, I argue against this objection. The core mistake, I suggest, lies in an impoverished conception of the context of use in which an ambient pattern achieves its informational significance. My strategy is threefold. First, I show that the objection tacitly relies on an implausibly narrow understanding of the context of use. Second, I develop the notion of coordination as offering a more adequate account of such contexts—one that makes room for the counterfactual robustness the objection demands. Third, I clarify the corresponding notion of specification. As a result, a proper modal view of usage-based information will emerge.

To begin with, the objection treats information based on use as exhausted by the very circumstances in which an ambient pattern is actually exploited here and now. But this is highly implausible. Ordinary talk of information-use is already counterfactual in character: to say that a pattern is used to perform a certain action is, at least in part, to say that the same pattern would remain usable for that action across the relevant range of circumstances, so long as certain background conditions obtain. If information is constituted in use, then the "context of use" cannot be identified with a single occurrent episode. Any adequate account must appeal to something that reaches beyond the particular situation in which the pattern is actually used. The objection's mistake is to take this requirement to be one the usage-based view cannot satisfy.

By contrast, I suggest that the notion of coordination provides a more illuminating picture of how information is constituted in use. Van Dijk & Kiverstein (2021) propose that ambient patterns become proper information insofar as they are used to achieve coordination with the environment. The crucial point is that coordination is not a momentary event, but a practice: it is not exhausted by any particular episode of successful engagement. Rather, it is essentially future-oriented. What unifies particular performances as instances of coordination is the agent's practical commitment to continue acting in certain ways under relevantly similar circumstances. This future-oriented character helps explain how possible uses are grounded in actual and past uses. The claim is that the opportunity for action—i.e., the affordance—associated with an ambient pattern is a projection from the history of established uses of that pattern in coordination (see Goodman, 1983). In other words, although a given pattern may be compatible with many possible actions in virtue of its intrinsic features, which possibilities are projected depends on the organism's history of interaction with that pattern. Past success in coordination entrenches some projections rather than others. In this way, the history of actual use for coordination constrains which affordance is projectible from a given ambient pattern. The relevant "context of use" that fixes informational significance is therefore not a momentary situation, but the temporally extended, future-directed practice of coordination. Once this is in place, the objection's conclusion no longer follows. If coordination is the proper context of use, then information based on use can specify opportunities

for action while still being counterfactually supported. Specification supervenes on this dynamic of projection: it remains a relation between ambient patterns and the animal's activities; but those activities should not be identified with whatever the agent happens to be doing at a given moment. Specification is itself modal: grounded in a history of interactions with the pattern, an ambient structure can specify certain opportunities for action even when it is not currently being exploited. Specification can thus range across different spatiotemporal contexts without thereby ceasing to be determined by actual use.

Tuesday 30th June 17:00 — Communication (Grote zolder)

Maria Polychronidou, Robert Hartsuiker, Petra Hendriks & Simone Sprenger: *Idiom Processing Across the Adult Lifespan*

As people age, language processing undergoes various changes that reflect both individual cognitive trajectories and cumulative language experience. On the one hand, age-related declines in domain-general cognitive functions, such as working memory capacity and inhibitory control, can increase the demands of online sentence comprehension and lexical access (Salthouse, 1996; Hasher & Zacks, 1988; Wingfield & Grossman, 2006). On the other hand, older adults possess a rich reservoir of semantic knowledge and pragmatic competence acquired through lifelong exposure to language, which can facilitate predictive processing (Kavé & Halamish, 2015; Verhaeghen, 2003; Umanath & Marsh, 2014). Rather than reflecting a uniform decline, aging entails a dynamic process in which older adults draw on crystallized linguistic knowledge and compensatory strategies to preserve, and occasionally enhance, language comprehension across the lifespan.

One domain where age-related changes are especially evident is the resolution of linguistic ambiguity, which requires readers to select the intended meaning among multiple alternatives. Constraint-based models argue that language users resolve such ambiguity by integrating probabilistic cues, such as word frequency, syntactic structure, and contextual support, in real time (MacDonald et al., 1994; Kuperberg & Jaeger, 2016). Although older adults typically retain strong language knowledge, age-related declines in executive functions can make it more difficult to efficiently integrate multiple cues during online processing (Dagerman et al., 2006; Caplan & Waters, 2005). Consequently, older adults tend to rely more heavily on highly constraining contextual information as a compensatory strategy to overcome reduced cognitive control (la Roi et al., 2020; Federmeier & Kutas, 2005).

Idioms offer a particularly informative case for studying how aging affects language comprehension, because they are ambiguous between a literal and a figurative interpretation (e.g., "spill the beans"). Comprehending an idiom involves accessing a stored figurative representation and suppressing the literal interpretation, especially in neutral contexts (Cacciari & Tabossi, 1988; Titone & Connine, 1999). While idioms are stored as holistic phrasal units (Sprenger et al., 2006) and idiom knowledge tends to remain robust or even improve with age (Carrol, 2023; Sprenger et al., 2019), their processing still requires executive control (Rommers et al., 2013). Recent eye-tracking research shows that older adults exhibit increased reading times when idioms appear in noncanonical structures, indicating reduced cognitive flexibility and a stronger reliance on familiar forms (Haeuser et al., 2021). Moreover, older adults appear to depend more on supportive context to facilitate idiom interpretation, compensating for declines in inhibitory control and working memory (la Roi et al., 2020).

In the present study, we investigated how idiomatic processing during reading evolves across the adult lifespan by combining eye-tracking with behavioural assessments of cognitive functioning. A total of 110 Dutch-speaking adults aged 30–80 years participated. All participants completed a sentence reading task while their eye movements were recorded, as well as an idiom familiarity

questionnaire and a battery of tasks assessing working memory, inhibitory control, cognitive flexibility, associative learning ability (Hake et al., 2023), and reading exposure using the Dutch Author Recognition Test (Brysbaert et al., 2020). The experimental materials comprised 44 Dutch idiomatic phrases paired with closely matched literal counterparts, all matched for syntactic structure to enable direct comparison, and each sentence was embedded in either a biasing context that supported an idiomatic interpretation or a neutral context lacking such cues, yielding four conditions: idiomatic in supportive context, idiomatic in neutral context, literal in supportive context, and literal in neutral context.

A Latin-square counterbalancing design ensured that each participant encountered every item in only one condition. Each trial presented the context sentence first, and the critical sentences followed a consistent subject–object–verb structure across all conditions with an identical initial noun phrase in the idiomatic and literal versions to prevent premature anticipation of the intended interpretation. In addition, 44 garden-path sentences and their controls were included as fillers to vary syntactic complexity and reduce predictability. Eye-tracking analyses targeted multiple measures in the target region as well as pre- and post-target regions to capture both early and later stages of processing. Data were analysed using mixed-effects regression models including fixed effects of idiomaticity, context predictability, age, and individual-differences measures, as well as random effects for participants and items.

The results showed that idiomatic expressions were processed more rapidly than their matched literal counterparts in both early and late eye-movement measures, indicating a reliable processing advantage for idioms. Context strongly modulated this pattern: in idiom-biasing contexts, idioms were processed more efficiently whereas literal continuations incurred increased reanalysis costs, while these effects were weaker in neutral contexts. Analyses of age did not reveal a global age-related slowdown. Instead, age-related effects were primarily observed in reanalysis measures, following a nonlinear pattern. To characterize individual differences, we conducted a principal component analysis on the cognitive measures, which yielded two factors corresponding to working memory and cognitive flexibility.

These cognitive factors, rather than chronological age per se, accounted for substantial variability in idiom processing. Higher working memory was associated with a larger processing advantage for idioms in both early and later measures, and cognitive flexibility selectively modulated the integration of contextual information and the suppression of literal interpretations. In addition, reading exposure and associative learning ability primarily affected later processing stages, strengthening context use and reanalysis patterns, while greater idiom familiarity reduced overall reading times. Together, these findings show that idiom comprehension remains efficient across adulthood, that supportive context guides processing and makes literal interpretations costly, and that individual differences in working memory and cognitive flexibility selectively support comprehension when contextual integration or suppression of literal meanings increases processing demands, rather than producing a general facilitation of reading.

Niklas Dahl: *Trust and the Chain of Communication*

There is a conflict between two intuitions in the debate on reference. On the one hand, as has been convincingly argued by Putnam (1973), Kripke (1981), and Kaplan (1989), speakers of a language seem to be able to refer to things which they do not themselves have discriminating information about. To paraphrase Kaplan's (1989) description of semantic consumerism, one does not need to know where Samarkand is to communicate referentially about Samarkand. There are also good reasons for believing that one does not need to know any description which picks it out in order to refer to it. On the other hand, we do also have countervailing reasons to believe that referential communication does actually require us to have discriminating knowledge

of referents to be successful. As Evans (1982) argues, to have a properly object-directed thought, one would need to have enough information to discriminate that object from other candidate referents. Hence, such thoughts can only be conveyed to hearers who are in a similar position to do so. And, more recently, Dahl (2025) has argued that to properly understand action-directing uses of referential expressions, one must have the ability to, in a sense, discriminate what the referential expression refers to (see Quine 1977 and Hawthorne & Manley 2012 for arguments to the contrary).

My aim in this talk is to reconcile, on the one hand, our ability to successfully communicate with referential terms whose reference we cannot independently discriminate with, on the other hand, the intuition that referential uptake requires us to know what is being referred to. I begin by considering the standard explanation of our ability to communicate referentially with terms which we do not have discriminating information for: an explanation in terms of the causal chain of communication linking a referent with a speakers' utterances of the referential term. This explanation, as Evans (1982) argues, is not sufficient for having information about the referent. As such, the mere fact that a term stands in such a relation to an object would not in itself provide any explanation of our ability to interpersonally co-ordinate actions by uttering the term. For it to do that, we would have to be aware of that causal connection so that we could use it to guide our search for the referent.

But this very point, I will argue, is what shows us how to solve the problem. What we need to be able to use borrowed terms to co-ordinate action is that the relevant discriminating knowledge is available to us. To some extent, this is already implicit in how we think about knowledge-who. As Farkas (2016) argues, we can be said to know what a phone number is even if we would have to look at the contact list in our phone to find the answer. Further, when we learn a new referential term from another speaker, we trust that they could successfully communicate referentially with it. If we did not, then we could not intend – as the proponents a causal chain of communication would have it – to use it to refer as they do. And that trust allows us to partly rely on them for our ability to pick out what the term is referring to. It is trust, not causation, which makes up the links of the chain of communication.

The idea, then, is that we can offer a shared explanation of both intuitions. We can refer using borrowed vocabulary because we trust other speakers to be able to aid us in locating and discriminating the referent. And that we trust certain speakers to be able to aid us in identifying a referent is the discriminating information we need for both object-directed thought and to have our actions successfully targeted by the use of a referential expression. Further, this approach also answers a criticism levelled against the causal version of the chain of communication, originating in Sosa (1970, cf. Hawthorne & Manley 2012), namely that it cannot explain reference by reverse causal chain. These are cases where, say, I order a ship to be built and named The Mary Sue. Several years later, without ever having been in causal contact with the ship, I can still refer to it by name. But whereas the connection goes in the wrong direction for the causal version of the chain of communication, an account where the links are made up by relations of trust can extend both forwards and backwards in time. I can trust that the agents sent out to construct and name the ship have acquired discriminating information they could provide me with to locate the ship.

Corijn van Mazijk: *Coordination without Meaning*

Archaeological and cultural-evolutionary debates about prehistoric symbolism remain dominated by a semiotic paradigm. Ornaments, pigments, figurines, and other non-utilitarian materials are typically interpreted as symbols: signs whose social significance derives from shared conventions and encoded meanings, often taken to parallel the emergence of language. This framework continues to structure discussions across archaeology, philosophy, and psychology, where

material culture is treated as external evidence for representational capacities such as reference, abstraction, and explicit meaning attribution. This paper argues that this symbolic framing is poorly suited to the prehistoric record. Many material practices that are routinely labeled “symbolic” appear to have played important social roles without functioning as symbols in the semiotic sense. Their effectiveness did not depend on users explicitly representing what these materials meant, nor on shared codes comparable to linguistic conventions. Instead, their social force arose from participation in structured practices that coordinated behavior, stabilized expectations, and regulated interaction across individuals and groups.

To capture this pattern, I introduce the concept of opaque social instruments (OSIs). OSIs are culturally transmitted material practices that organize social relations while remaining largely opaque to the explicit reasoning of their users. They are instruments because they reliably perform social work—structuring affiliation, cooperation, exchange, and boundaries. They are opaque because participants typically do not represent these functions as such; instead, they experience the practices as customary, beautiful, prestigious, sacred, or simply obligatory. Social coordination is achieved without conceptual transparency. The OSI framework challenges a core assumption of the symbolism paradigm: that socially effective material culture must operate via symbolic encoding or transmitting meanings. Instead, OSIs rely on the way material forms become embedded in shared practices and learned through participation. Individuals do not need to know why a practice works in order for it to work. Cultural stability, not representational clarity, is the relevant explanandum.

The paper situates OSIs within a cultural evolutionary framework, drawing on well-established work showing that cultural traits persist and spread when they reliably exploit evolved learning biases and coordination pressures. From this perspective, prehistoric ornaments and related materials should not be treated as exceptions that demand symbolic explanation. Their long-term persistence strongly suggests that they, like rituals and institutions in later societies, solved recurrent social problems in ways that were cognitively tractable and culturally transmissible.

Six families of evolved psychological biases are particularly relevant for understanding how OSIs function and why their effects remain opaque. First, kinship and in-group favoritism biases incline individuals to preferentially trust and invest in perceived group members. Material markers can extend or reshape in-group boundaries, enabling cooperation beyond close kin without users explicitly representing this function. Second, coalition and alliance tracking mechanisms make humans highly sensitive to cues of group membership and social alignment. Visually salient material practices can function as stable coalition markers, supporting trust and coordination even when participants experience them merely as “what people like us wear or use.” Third, precaution and threat-management biases reflect long evolutionary histories of intergroup tension alongside the necessity of intergroup exchange for gene flow. OSIs can buffer interaction in risky social contexts—through gifting, ornamentation, or display—while being interpreted phenomenologically as custom or prestige rather than as conflict regulation. Fourth, prestige and mate-choice biases render individuals especially attentive to cues associated with skill, success, or social standing. Material practices tied to craftsmanship, hunting success, or long-distance exchange can organize sexual selection and status differentiation without being consciously understood as such. Fifth, conformity biases and sensitivity to opaque procedures lead humans to copy practices with high fidelity even when no one can articulate their rationale. This allows socially effective practices to persist over generations while remaining experientially opaque. Many material traditions are reproduced because “this is how it is done,” not because their social function is represented. Finally, narrative memory and minimally counterintuitive content enhance the transmission of certain material forms by anchoring them within stories, myths, or ritual frameworks. While these

narratives may provide local explanations, they often further obscure rather than reveal the underlying social functions of the practices they sustain.

Together, these biases explain how OSIs can be both socially powerful and cognitively opaque. Coordination is achieved through patterned participation, not through shared meanings in the linguistic or semiotic sense. Reframing prehistoric material culture as OSIs shifts the explanatory focus from representation to organization. Instead of asking what artifacts meant or what they symbolized, the OSI framework asks how material practices stabilized coordination, structured social relations, and persisted across populations. This reorientation aligns archaeology more closely with philosophy of psychology and cultural evolution, offering a model of extended social cognition in which material forms actively organize social life without requiring meaning attribution.

Tuesday 30th June 17:00 — Structure of Representation (Spiegelzaal)

Marko Jurjako: *Personal and Subpersonal Explanation: The Interface Problem Revisited*

In the philosophy of cognitive science, a prominent distinction is drawn between personal and subpersonal levels of explanation. However, it is still an open question how these levels are related. This question is often described as the interface problem. There are different approaches to resolving the interface problem: autonomism, functionalism, and the co-evolutionary model. In her influential work, however, Zoe Drayson treats the interface problem as specific to the functionalist framework, while regarding autonomist approaches as not being about the personal/subpersonal distinction at all.

I argue against this restrictive understanding of the interface problem and maintain that the relation between personal and subpersonal explanations can be understood in a plurality of ways. I support this view by revisiting Daniel Dennett's original introduction of the personal/subpersonal distinction. By examining Dennett's discussion of pain, reason-based action, and consciousness, I show that personal-level constructs may be: 1) autonomous from subpersonal constructs; 2) functionally mapped onto them; 3) or revised in light of them. Based on these considerations, and against Drayson, I defend a pluralist understanding of the interface problem, according to which functionalist and autonomist approaches should be seen as competing solutions, rather than as reflecting incompatible conceptions of the personal/subpersonal distinction itself.

Yinzhu Yang: *Analog and Digital Representation Reconsidered: A Network-Based Approach*

Where, if anywhere, is the border between perception and cognition? A family of answers appeals to representational format: perceptual states are said to be iconic, while cognitive states are discursive or language-like. Yet the contemporary landscape is crowded with borderline phenomena that seem to possess both perceptual and cognitive markers. I begin with two pressure points. In billiard-ball causation, we have experiences that display perceptual signatures like immediacy, adaptation, and known-illusion persistence, while also trainable and responsive to background expectations. Core cognition cases in general raise similar worries: the same system can appear perceptual in its automaticity and cue-dependence, yet cognitive in its integration with reasoning and learning.

In response to these problems, I suggest we need to go back and reexamine the concept of representational format, which I believe is the culprit to blame for our assumption of a hard and fixed boundary between cognition and perception. I argue that these disputes are partly sustained by a tacit Rigid Format Assumption: analog and digital formats are mutually exclusive, sharply bounded natural kinds. Against this, I motivate a Soft Format Assumption by emphasizing hybrid analog–digital systems, both in engineering and in neural computation, that require interface components for converting and coordinating representational vehicles. Once hybrid interfaces are

taken seriously, it becomes less plausible that “analog” and “digital” neatly partition whole psychological systems such as perception and cognition. My central proposal is that representational format should be analyzed one explanatory level down, as properties of network substructures. The concept of network plays an important role in my formulation of analog and digital representation. My purpose is to argue that the distinction between formats can be flexible because their underlying network is flexible. I propose that analog and digital representations are better understood in terms of properties of network science, instead of immutable, fixed kinds like circles and squares. This characterization has many benefits: In particular, it would help us form a novel understanding of many problems in cognitive science, including the debates regarding cognition-perception border and cognitive ontology. As is mentioned above, current accounts of the cognition-perception border face serious challenges, such as the problem of core cognition and the problem of higher-level perception. A network account could dissolve the rigid border by characterizing cognition and perception in a more dynamic way, thus supporting The Soft Format Assumption. Within a complex computational network, analog representations correspond to single nodes or sub-networks whose quantitative state tracks represented values. Their representational role can often be interpreted at the level of a variable. Digital representations correspond to compound subgraphs whose organization determines what discrete token combinations mean. Their representational role depends essentially on connectivity patterns, compositional relations, and decoding rules distributed across multiple units. Crucially, genuinely hybrid systems require interface nodes: components that translate, discretize, or transform signals so that information can pass between analog-like tracking and digital-like token manipulation. Format differences, on this view, are not metaphysical kinds but explanatory properties of how information is organized and transformed in the network. I then show how the network framework reframes the motivating borderline cases.

In short, this paper undermines the expectation that a single, rigid format boundary can do the explanatory work demanded of the perception–cognition distinction.I propose that the abovementioned problems could be overcome if we view digital and analog representations in the lens of network analysis. The interactive network view handles the “hard cases” like core cognition and higher-level perception with ease. They are expected products of interactive subgraphs with layered formats. An analog magnitude representation feeding into a conceptual node, or a visual object file linked to a memory index, is exactly the kind of structure a network model predicts. By allowing hybrid representations and interaction, this framework reflects the mind’s true complexity. It explains how we can perceive a number or a substance kind directly, why infants can reason about objects before language, and how conceptual thinking can utilize sensory-like simulations without breaking the unity of the cognitive system. This is both empirically plausible and theoretically satisfying, since it replaces an unnecessarily strict format boundary with a fluid account of how perceptual and conceptual information could interact.

Yağmur Deniz Kısa, Roman Stengelin, Luke Maurits and Daniel B.M. Haun: *Cognitive Change Without Linguistic Change: The Rise of Egocentric Frames of Reference in the Hai||om*

A long Western tradition in philosophy, psychology, and neuroscience has assumed a cognitive universal: humans think about space primarily egocentrically, relative to the left, right, front, and back of their own bodies. This assumption has been challenged by substantial cross-cultural variation in spatial cognition and language: Preferring to use egocentric frames of reference to talk about space seems to be limited to globalized, industrialized societies or groups heavily influenced by such societies. Outside such groups, many indigenous languages prefer geocentric frames of reference, relying on cardinal directions (east/west/north/south) or environmental features (uphill/downhill) to describe spatial relations. Speakers of egocentric and geocentric languages not only prefer to talk, but also prefer to think egocentrically and geocentrically,

respectively–revealing substantial cognitive diversity in a fundamental domain of cognition. Where does this cognitive diversity come from? So far, the dominant explanation has been the Whorfian hypothesis: Preferring to talk egocentrically causes people to also think egocentrically, even when not using language. If this is true, then a population that is having a shift in their frames of reference should have a linguistic shift before a cognitive shift, but not vice versa.

Here, we provide evidence suggesting an ongoing shift in spatial cognition—but not language—within a culture, showing that Whorfian views are insufficient on their own to explain cognitive diversity in spatial frames of reference. Hai||om people from rural Namibia have consistently been shown to prefer geocentric FoR both in language and cognition (Neumann & Widlok, 1996; Haun et al. 2006; Widlok, 2007; Haun & Rapold, 2009; Haun et al. 2011). In the present study, we returned to the same Hai||om village roughly two decades after the original research. Our goal was to conceptually replicate the robust pattern that Hai||om prefer to talk and think geocentrically, using the most standard and widely used frame of reference tasks and drawing on a wide, age-diverse sample of adults. In Study 1, 30 Hai||om participants completed two tasks. First, in the cognitive, animals-in-a-row task, participants studied an array of animals and then were asked to reconstruct the array from memory after 90 degree rotation. There are at least two correct solutions for what counts as the same. If participants memorized the animals egocentrically, as all facing left, they should reproduce them facing left after rotation; if they memorized them geocentrically, as all facing east, they should reproduce them all facing east (Fig. 1a). Second, in the linguistic, director-matcher task, the director was asked to describe simple scenes (i.e. different spatial relationships between a toy bucket and a toy dog) to the matcher so that the matcher can build the same scene (Fig. 1c). The scenes can be described using egocentric (e.g. “the bucket is to the left of the dog”), geocentric (e.g. “the bucket is to the east/sunrise of the dog”), mixed, or object-centered language (e.g. “the bucket is behind the dog”). Much to our surprise, in the cognitive task, Hai||om participants showed a strong preference for egocentric frames of reference: 28 of 30 participants (93%) rearranged the animals egocentrically after rotation. This contrasts sharply with earlier data from the same community collected in 2005 (Haun et al., 2011), in which 100% of trials showed a geocentric preference and no egocentric responses (Fig. 1b). We fit a Bayesian categorical regression model both to the present and the historical data.

Our results revealed very strong evidence for an egocentric shift in Hai||om: The 95% HPD interval for the difference in the probability of egocentric outcomes when comparing the current data to the historical data excluded zero (0.66-0.84) and indicated a mean increase of 0.75. However, Hai||om speakers showed no such egocentric shift in language use: In the director-matcher task, they described spatial relationships among objects primarily using geocentric strategies, just like they did in the past (Neumann & Widlok, 1996; Fig. 1d). Critically, unlike what we observed in the cognitive task, we did not observe a greater preference for egocentric FoR in language: the 95% HPD interval for the difference in the probability of egocentric outcomes when comparing the current to the historical data did not exclude zero (-0.22-0.09) and had a negative posterior mean (-0.05). Using close adaptations of the original task, we observed different cognitive preferences among the Hai||om across two time points, suggesting a change in spatial cognition over time.

However, this difference may reflect sampling or methodological variation rather than historical change. Haun et al. (2011) tested children aged 7–11, whereas we tested adults aged 23–66. In addition, although our task closely matched the original, small procedural differences existed and any of these differences could, in principle, account for the divergent results. In Study 2, conducted two years after study 1, we ran a direct replication of the cognitive task by Haun and colleagues (2011) to rule out explanations based on measurement changes or sampling differences across the two time points. Testing both age-matched participants (children ages 7-11) and retesting the original participants now as adults two decades later in the same paradigm, we found robust evidence that Hai||om show a greater preference for egocentric FoR compared

to the past (Fig. 2). Comparing new data to historical data from the same community, we found that contemporary Hai||om show a greater preference for egocentric frames of reference compared to previous reports, suggesting an egocentric shift in their cognition. Surprisingly, we document no such shift in their language.

Together, these results show that cognitive shifts in spatial reference frames can occur without parallel changes in language. Non-linguistic factors appear to be at play in promoting the rise of egocentric thinking in Hai||om people. We suggest that increased exposure to egocentric material culture (e.g. cars, books, screens) from urban centers may foster egocentric thinking by making egocentric distinctions behaviorally relevant and by representing space egocentrically. Material culture, rather than language, may therefore be a key driver of diversity in spatial thought.

Tuesday 30th June 17:00 — Reasoning & Attention (Voorkamer)

Ying Nortrup: *When Boredom Sustains Attention*

Contemporary accounts increasingly characterize boredom as a regulatory signal that motivates agents to disengage from unsatisfactory cognitive engagement and seek more rewarding alternatives. Yet a familiar feature of digital life appears to challenge this view: individuals often turn to social media when bored and remain engaged in prolonged scrolling even when the experience itself remains unsatisfactory. This phenomenon can be formulated as a paradox for the functional theory of boredom. If boredom functions to terminate unsatisfactory engagement, sustained scrolling should be unlikely; yet boredom frequently appears to sustain it.

Drawing on theories of attentional bias and opportunity-cost evaluation, four responses to this paradox are examined. Digitally structured environments can sustain exploratory attention even when individual stimuli remain unsatisfying, revealing a structural limitation that needs to be expanded in the functional theory of boredom.

Carla Sebastián Enesco, Nerea Amezcua-Valmala & Federica Amici: *Waiting for a better reward: Anticipatory & regulatory strategies in parrot delay of gratification*

The ability to delay gratification (DoG) is commonly taken as a core indicator of self-control and future-oriented decision-making. In both human and non-human cognition, DoG performance is typically operationalized by the maximum waiting time for a delayed reward or by success across progressively increasing delay intervals. More recent work, mainly in developmental psychology, has highlighted the importance of examining what individuals do during the waiting period itself, arguing that spontaneous self-regulatory strategies deployed while waiting can inform not only whether individuals succeed but also how individuals succeed (Neuenschwander & Blair, 2017). Behaviors observed during waiting are usually distinguished according to their functional relation to the delayed incentive. Some behaviors remain directly oriented toward the reward, such as sustained attention to or attempts to access it and are often described as anticipatory or motivational strategies. Other behaviors are not directed at the incentive and instead appear more plausibly related to sustaining waiting by modulating arousal or attention. These behaviors are typically described as regulatory or volitional strategies, for instance through engagement in alternative activities (Duckworth & Steinberg, 2015; Metcalfe & Mischel, 1999). Importantly, beyond this functional heterogeneity, findings from human studies suggest that success in delay tasks is not uniquely associated with the deployment of a particular strategy. Instead, it seems to depend on how motivational and regulatory strategies interact during the waiting period

(Neuenschwander & Blair, 2017). Taken together, these findings challenge a unidimensional view of self-regulation and provide a more nuanced understanding of delay behavior.

Despite these advances, this strategy-based approach to self-control has been explored almost exclusively in humans, leaving open the question of whether these patterns are specific to our species or instead reflect more general features of delayed decision-making. The present study is situated within this broader question. We focus on parrots, a taxon that, despite its phylogenetic distance from humans, shows behavioral and cognitive capacities that in several respects resemble those described in primates (Lambert et al., 2018). The still limited body of research suggests that parrots perform well in domains central to self-control research, including problem solving, future-oriented behavior, and behavioral flexibility (Rössler & Auersperg, 2023). Rather than focusing solely on whether parrots are able to wait, the present study examines how waiting is sustained. To this end, we extend a strategy-based analysis of delay behavior to a non-human species, asking whether systematic patterns in waiting behavior—such as differences between anticipatory and regulatory strategies, and their interactions—can be identified beyond the human lineage.

We tested 8 macaws (Ara spp.) in an intertemporal choice task adapted from established DoG paradigms. In each trial, subjects could choose between an immediately available low-value food reward and a preferred high-value reward available after a delay. Delay intervals increased incrementally across sessions, and individuals advanced to longer delays only if they met predefined performance criteria. As in prior DoG research, performance was assessed in terms of success across increasing delays, and we additionally examined patterns of early abandonment (“giving up”), defined as consumption of the immediate option before reaching the maximum tested delay. Beyond these outcome measures, we analyzed subjects’ behavior during the waiting period using a newly developed ethogram designed to capture functionally distinct coping strategies. Specifically, we distinguished three broad classes of behaviors: (i) reward-oriented behavior, which involves attentional, postural, or motor orientation towards the inaccessible delayed reward; (ii) motor self-regulation, encompassing a set of body activities ranging from localized movements (head, body) to pacing around; and (iii) self-maintenance, including all kind of self-referential behaviors (e.g., preening, beak cleaning). For reward-oriented behavior and motor self-regulation, behaviors were further organized into hierarchical levels of increasing intensity, allowing us to examine not only whether particular strategies were deployed, but also how their intensity varied as task demands increased.

Overall performance followed a sigmoidal curve, with high success up to a delay threshold followed by an abrupt drop—rather than a progressive discounting. This pattern mirrors recent accounts in humans and nonhuman primates showing that the subjective value of a delayed reward remains relatively stable until a temporal boundary is reached, after which persistence collapses sharply (Green & Myerson, 2004). Notably, when individuals gave up waiting, they typically did so early in the trial, suggesting a rapid, categorical decision about whether to persist rather than a gradual erosion of self-control over time.

Several patterns emerged from the behavioral analyses. Increasing delay demands were not associated with a uniform increase across all classes of coping behavior. Motor self-regulation showed a systematic shift toward higher-intensity forms as delays increased, suggesting a growing reliance on bodily regulation during prolonged waiting. In contrast, orientation toward the delayed reward did not show a monotonic increase with longer delays: lower-intensity orientation (e.g, facing or pointing toward the reward) remained relatively stable across delay conditions, higher-intensity reward-directed behaviors (e.g., active pecking) increased at intermediate delays

but dropped sharply at longer delays. Self-directed behaviors were comparatively infrequent and remained stable across delays.

Taken together, these findings suggest that as waiting demands increase, individuals rely more on strategies plausibly related to arousal modulation, while sustained anticipatory engagement with the delayed reward appears to break down under high delay demands. This pattern mirrors developmental findings showing that delay behavior is supported by qualitatively different strategies that respond differently to increasing temporal demands, rather than by a single, uniformly increasing form of self-regulation.

Ignasi Gil Gómez: *The Epistemic and Zetetic Irrationality of Incessant Checking in OCD*

Consider the following case:

Do I have HIV? Stella sits on the edge of her bed, the room still and silent except for the faint hum of her phone charging. She knows she used a condom that night, four months ago—she remembers checking twice—but the thought won't stop looping. What if it broke? What if it slipped? She's taken nine HIV tests since then, each one negative, each followed by a brief calm that dissolves within days. What if the tests were wrong because there was a mistake in the lab? What if, despite all that is known about the window period, the virus is still hiding, waiting? What if someone is sabotaging the tests so she can never know if she's infected? Stella keeps taking HIV test after HIV test, searching for a certainty she never finds.

Incessant checking is one of the most common compulsive behaviors used to relieve the anxiety caused by obsessions in OCD (APA, 2013). It is widely agreed that Stella's behavior can be practically irrational, as it only reduces her anxiety temporarily and may interfere with her daily goals. For instance, she has begun cancelling plans with friends and missing work deadlines because she spends hours researching testing accuracy or going out to take another HIV test. And the relief she gets from each negative result never lasts: it fades within days, sometimes hours, leaving the original distress intact. In this way, each new test does not resolve her anxiety but reinforces it—she learns that the next time the worry returns, the only way to cope will be to check again.

But what about her irrationality as an inquirer? Is there something irrational in what Stella believes about her inquiry, or in the way she carries it out? In this paper, I take a closer look at these questions. I argue that we must distinguish between two scenarios: in one, she keeps inquiring because she holds incorrect beliefs about her prospects for success in inquiry. In the other, she does so despite holding correct beliefs about those prospects. In the first case, Stella's failure is epistemic, while in the latter case it occurs at the level of inquiry (i.e., the failure is zetetic). In other words, incessant checking in OCD can be either epistemically or zetetically irrational, depending on why Stella violates the SUCCESS NORM OF INQUIRY, a version of which has recently been defended by Hubacher Haerle (forthcoming).

SUCCESS NORM OF INQUIRY: S ought not to inquire whether P? at time t if, at t, it is irrational either to believe that one will achieve the aim of the inquiry or to suspend judgment about it.

Distinguishing between these two forms of irrationality allows for a finer-grained understanding of incessant checking in OCD than accounts that treat it as uniquely epistemically irrational (see Friedman, 2019) or zetetically irrational (see Hubacher Haerle, 2023, forthcoming). It also brings

into view a dimension of insight that is not captured by standard clinical assessments, namely, whether the agent correctly evaluates her prospects for success in inquiry.

Tuesday 30th June 17:00 — Agency (Bovenkamer)

Kaisa Kärki & Michael Laakasuo: *The Fundamental Skills Approach to Deskilling*

One central aim of automation has been to develop machines that can perform task previously carried out by humans. While there seems to be nothing wrong with using a robot vacuum cleaner or a pocket calculator, there seems to be something wrong with relying on a large language model when choosing which candidate to date or when writing an academic essay for the first time. In this paper, we present a theory that explains where the difference lies – when we should avoid deskilling, that is, the loss of skills by outsourcing tasks to artificial intelligence. We draw from a concurrent discussion on the aims of education in philosophy of education arguing that deskilling should be avoided in what we call fundamental skills; those tasks, skills, and capacities that construct individual or collective human autonomy.

In our view, they are (1) general capacities for making informed decisions over the course of a person's life, (2) skills and capacities for reaching and maintaining democratic decision making in human communities, and (3) skills and capacities that cannot be simulated by artificial intelligence in the first place – such as the capacity for moral cognition and the capacity for value-based decision making. We present four arguments for this view: an argument from unhealthy dependence, the case of an electromagnetic pulse, an argument from the human value of autonomy, and an argument from the right to open future. Then we answer three objections. The upshot of our approach is that it helps us identify choices in which people should refrain from consulting large language models and determine those educational settings in which the use of artificial intelligence should be discouraged.

Polaris Koi: *Options as mentalia*

When I make a choice, I am faced with a set of options. But how come my options are the ones they are? And what does it take for something to be an option? Despite being often overlooked, the generation of a decision set is a necessary feature of intentional action. The nature of decision sets is of relevance across diverse areas including but not limited to philosophy of action, work on moral responsibility, and philosophy of economics. However, in theorising about intending, choosing, and acting, the presence of a salient, well-defined decision set is a starting point rather than the topic of inquiry. Philosophical work on options and decision set generation is scarce (though see Kalis, Kaiser, & Mojzich 2013; Smith 2010; Hedden 2012; Jeppsson 2018; Chappell 2008), and most of it is normative. To remedy this, this paper develops a systematic, descriptive account of options and their generation.

In what follows, I take an agent to be an entity with the capacity for intentional action or, at minimum, for endogenous selection among competing courses of behaviour. I take a decision set to be a set of items that an agent chooses from. Rather than cumbersomely speak of decision set constituents (DSCs), I will refer to DSCs as options.

On a view that is widespread across the decision sciences, options are simply any alternatives or behavioural trajectories available for an agent in the world. This is an influential approach taken as a premise across multiple literatures: for example, the notion of overall freedom (Pattanaik & Xu 1990) is associated with the number of alternatives available for the agent in the world. In behavioural economics (Glimcher 2011), with its popular notion of ‘nudging’, alternatives are thought of as being ‘out there’ in the world (Thaler & Sunstein 2008) rather than contingent on the agent’s mental life. The notion of revealed preferences, which is widespread in economics, also

takes options – understood as the objects of preference and choice – to be out in the world (although see Thoma 2021 for an unusual, contrary view). For example, when an economist infers that readers prefer best-selling books over those that do not sell, they take the agents' decision sets to include all the books being compared.

Conceiving of options as mind-independent and agent-external makes for very large sets of options. For example, someone choosing where to take their partner for dinner has all the restaurants in town in their decision set. There are, however, several reasons to believe that decision sets are smaller than the range objective alternatives, unless the set of such alternatives is very small to begin with. The most obvious of these is epistemic: the person choosing the restaurant might not know all the restaurants in town. As such, the sets of items that agents, in fact, decide among must be in some sense subjective.

There are two ways that that we might construe of subjective decision sets. a. Subjective decision sets are made up of 'external options'. The set of external options is winnowed down by some mechanism(s) contingent on the agent's mental life, such as epistemic constraints (Caplin & Dean 2015; Smith 2010) or 'sensitivity mechanisms' (McClelland 2020). b. Embrace mentalism about options. Options themselves are mind-dependent. Subjective decision sets are made up of options as mental content, specifically, as mental representations of courses of behaviour.

Alternative A, however, does not fully capture the scope of options agents decide among. Options are sometimes products of our imagination, misapprehension, recollection, or creative synthesis. A painter deliberating on what to create is not selecting from things that are found in the world as such. Rather, she is envisioning something new. Such options are best understood as mentalia, rather than as something discovered in the world. Rather than taking some options to be mind-independent possibilities and others to be mental representations, I will offer a unified account on which all options are mentalia. This allows us to consider options as mental states and to apply what we know about the cognitive processes associated with mental representations (see, e.g., Johnson-Laird 1983; Smortchkova, Schlicht, & Dolega 2020) to understanding decision sets and their constituents.

On the positive account developed here, i.e., mentalism about options, options are construed as mental representations, that is, as mental states with semantic content. However, not all mental representations are options. Instead: Options are mental representations of courses of behaviour with first-person guidance that agents take to be feasible for themselves.

If we construe of options as mental representations, then option sets are finite. This contrasts with the infinite range of actional opportunities afforded by the mind-external world. Since agents must generate a finite set of representations (one that fits their cognitive bandwidth), mentalism about options gives us a naturalistic account of the finitude of decision sets.

Our past and present circumstances influence option generation through multiple mechanisms: through the availability of information, and by modulating agents' conceptions of the world and of themselves, including what agents take their own capabilities to be. Option mentalism predicts that if agents can generate a mental representation of a given behavioural trajectory, but that trajectory does not emerge as an option, then the option omission results from a negative feasibility assessment. This allows us to disambiguate two ways in which individual background factors, such as disadvantage, shape choice. Even in the presence of salient information that allows someone to imagine becoming a doctor, their self-model might forestall the emergence of this behavioural trajectory as accompanied by an ascription of feasibility, and therefore preclude

it as an option. By contrast, if a behavioural trajectory cannot be imagined, then option omission results from the limitations of the information that the agent is operating with.

The account of options as mentalia offered here captures the subjective, agentive nature of decision sets, elegantly explains how choice from a manageable set of options is possible in complex environments, and yields a conceptual basis against which future work on decision and choice can be developed.

Jonida Kodra: *Control and Sense of Agency in Auditory Verbal Hallucinations*

Auditory verbal hallucinations (AVHs), broadly defined as experiences of hearing voices in the absence of an actual speaker (APA, 2022; David, 2004; Slade & Bentall, 1988), are phenomena common in both clinical and non-clinical populations (de Leede-Smith & Barkus, 2013; Baumeister, 2017). AVHs are clinically and philosophically important because they raise key questions about the nature of hallucinatory experiences, mental agency, and the formation of certain delusional beliefs (Connors et al., 2016; Stephens & Graham, 2000). I propose that a key aspect of AVH experiences that can help answer some of these questions and overall more clearly understand this phenomenon is control. Because the clinical literature has conceptualized control in heterogeneous ways (Swyer & Powers, 2020; Kern et al., 2015; Haddock et al., 1998), a more specific suggestion is to focus on control conceptualized as an ability and the exercise of the ability to be in control. Control so-understood already features prominently in three important debates. One such debate is on how to define hallucinatory experiences, including AVHs. Most diagnostic manuals and clinicians which consider control to play an important role within their definitions of hallucinations, tend to characterize AVH as a phenomenon over which voice-hearers have no control (APA, 2022; Slade & Bentall, 1988). A second debate is about the nature of hallucinatory experiences. Some researchers have taken empirical findings that voice-hearers lack control over AVHs to indicate that AVHs are non-veridical perceptual as opposed to thought-like experiences (Aleman & de Haan, 1998). Specifically, those who distinguish between AVHs that are subjectively indistinguishable from veridical auditory perceptions and AVHs that are not, maintain that it is the lack of control that differentiates subjectively distinguishable AVHs from other inner experiences such as auditory mental imagery (Farkas, 2012; Knappik et al., 2022). The third debate relates to the etiology of AVHs. Some theorists argue that the reason why voice-hearers experience internally-generated phenomena (e.g., memory, imagination) as AVHs is also because they lack control over the onset, offset, and all features of the content of such episodes (Wu, 2012); others propose different causes for AVH experiences, but still posit that voice-hearers cannot voluntarily initiate AVH episodes (Jones & Fernyhough, 2007; Langland-Hassan, 2016).

In response, I want to make two suggestions. First, I suggest that a more careful consideration of the clinical literature on control shows that empirical findings are in contradiction with the views I just presented. Clinical data show indeed that voice-hearers possess the ability to be in control. More specifically, empirical findings suggest that when exercising control, voice-hearers want and are able to generate, inhibit, and change some of the features of AVH episodes (Moritz & Larøi, 2008; Nayani & David, 1996). Against the abovementioned views (Aleman & de Haan, 1998; APA, 2022; Slade & Bentall, 1988; Farkas, 2012; Knappik et al., 2022; Jones & Fernyhough, 2007; Langland-Hassan, 2016; Wu, 2012), and in line with both David's (2004) suggestion and empirical findings, I propose that voice-hearers lack the feeling or experience of control, but not the ability. Specifically, the suggestion is not that all AVH episodes are under the control of voice-hearers. The claim is rather that voice-hearers occasionally voluntarily generate, inhibit, and change some

of the features of their AVH episodes. Nonetheless, despite the exercise of their ability to be in control, voice-hearers still lack the feeling of having done so.

My second suggestion is that, in addition to disconfirming some views, empirical data can help to answer some important philosophical and clinical questions. For one, they allow for a clearer understanding of the process underlying AVH experiences and, by extension, also of the process that allows one to experience one's thoughts or thought-like experiences as self-generated, standardly known in the clinical and philosophical literature as sense of agency (Gallagher, 2004; Braun et al., 2018). A suggestion commonly put forward by several clinicians and philosophers in fact is that AVHs are internally-generated phenomena that are experienced as non-self-generated due to a disrupted sense of agency (Jones & Fernyhough, 2007; Stephens & Graham, 2000). In light of the proposals that voice-hearers have the ability to voluntarily generate AVH episodes and that AVHs are the result of an impaired sense of agency, a more specific suggestion is that empirical findings indicate that the process underlying the sense of agency might be one voice-hearers have some control over.

In addition, the ability and exercise of control over some of the features of AVH episodes (e.g., the content, volume, etc.) also indicate that the mental phenomenon that the voice-hearer experiences as non-self-generated is a phenomenon which one can normally exercise some control over (e.g., inner speech, imagination, etc.) (Gregory, 2016; Langland-Hassan, 2016), and not an auditory perceptual experience as suggested by other theorists. For another, empirical findings also seem to support the hypothesis that the exercise of control is not always accompanied by the feeling of exercising control (David, 2004).

Wednesday 1st July 09:00 — Keynote (Kerkzaal)

Mazviita Chirimuuta: *Evidencing Biological Naturalism*

In discussions concerning the possibility of AI consciousness, *computational functionalism* has been the dominant position. This is the view that consciousness is essentially a kind of computation and therefore medium (or substrate) independent (not reliant on any one kind of physical realizer). The opposing position, *biological naturalism,* has recently been discussed by Block (2025) and Seth (forthcoming). This is the view that consciousness is essentially a property of (some) living organisms and is therefore medium dependent. According to Block, the debate between computational functionalism and biological naturalism is over whether it is the computational *role* or the material *realizer* that determine consciousness. Here I argue that biological naturalism is better understood as the thesis that in living, conscious beings, roles and realizers are inextricably related in such a way that the computational functionalist requirement of medium independence cannot obtain. I discuss findings in neuroscience that are relevant to this issue. Although the neuroscience is still not settled, we can already rule out claims of medium independence of brain processes based on consideration of the greater energy efficiency of medium dependent processing.

Wednesday 1st July 10:45 — Plenary Symposium (Kerkzaal)

Federico Adolfi, Dimitri Coelho Mollo & Beate Krickel: *The Bases of Cognition: medium-(in)dependence, biological constraints, and the feasibility of computational-mechanistic explanation*

Beate Krickel: *Can Computational Explanations Be Mechanistic?*

Computational models are ubiquitous in cognitive neuroscience, yet their explanatory role remains unclear. New mechanists typically maintain that explanations are explanatory insofar as

they describe mechanisms. However, computational explanations seem difficult to accommodate within standard mechanistic accounts. In this short talk, I argue that two challenges arise: computational properties often fail to satisfy traditional criteria of constitutive relevance, and model-to-mechanism mappings alone do not explain why computational descriptions are explanatorily informative. Drawing on the case of collision avoidance in locusts, I propose a contrastive account of mechanistic explanation according to which computational models contribute by identifying difference-making features across possible mechanisms.

Dimitri Coelho Mollo: *Explanatory pluralism, cognitive ontologies, and medium-(in)dependence*

In this short talk, I will try and defend a pluralist approach to the discussion about the medium-(in)dependence of cognition and consciousness. I will suggest that it may be a mistake to think that there is a single answer to whether those capacities are medium-(in)dependent or else. In brief, medium-(in)dependent explanations are likely to be better suited for certain kinds of explanatory purposes and targets, in their turn better captured by specific choices of explanatory and target ontologies, some of which involving medium-dependent, some of which medium-independent individuation criteria.

Federico Adolfi: *What medium (in)dependence could mean for the feasibility of neurocognitive explanations*

In this short talk I will examine the epistemic consequences of medium (in)dependence assumptions, drawing on (applied) theoretical computer science. I will first discuss how computational complexity considerations can be brought to bear regardless of whether one rejects computationalism or indeed computation as a pragmatic perspective on cognitive systems. I will then explore the idea that medium (in)dependence plausibly affects the resource demands of obtaining certain epistemically useful objects, with consequences for the feasibility of neurocognitive explanations and knowledge discovery. Finally, I will argue that the complexity-theoretic interplay between medium (in)dependence and our explanatory goals has implications of broad interest that scientists and philosophers should explore together.

Parallel Sessions

Wednesday 1st July 14:30 — Symposium: Cognition in Action (Kerkzaal)

This symposium brings together cutting-edge theoretical and experimental work from philosophy and psychology, exploring how action plans and bodily constraints shape cognition. While most modern thinkers reject Cartesian dualism, its traces continue to linger in the study of brain and cognition. Cognitive brain function still tends to be studied in isolation from the rest of the body and the outside world, treating participants as passive receivers of stimuli and overlooking their role as goal-driven agents whose intentions and actions fundamentally shape how they filter, perceive, and use information. However, the brain does not operate in isolation: it is situated in a body, shaped and constrained by physiological needs, motor capabilities and the affordances for action in the world. Recognizing this is not just a philosophical refinement: adopting a 4E (embodied, embedded, enactive, extended) perspective fundamentally changes how we understand and experimentally study the brain and human cognition, as will become apparent through the talks in this symposium. Although 4E ideas are well developed philosophically, they have long remained at the periphery of mainstream empirical psychology. This symposium helps bridge this gap and illustrates the value of strong theoretical grounding in cognitive science. Using work from a range of fields and methods, it also offers methodological insights into how the role of action can be studied in cognition. This includes interactive paradigms that incorporate bodily

engagement, as well as strategies for adapting more traditional techniques, like fMRI, which does not permit movement, to address action-related questions through thoughtful experimental design.

In this symposium, Dr. Kiverstein will begin by making the case for an action-oriented cognitive ontology based on findings from cognitive and affective neuroscience. This is followed by a presentation by Dr. Groen of experimental findings from combined human neuroimaging (fMRI and EEG), behavior and AI modeling studies, which show that visual brain regions long thought to encode purely perceptual aspects of natural scenes represent their locomotor affordances (e.g., whether you can climb, swim or walk in the scene), contrary to commonly used, passively trained deep learning models of visual perception. Next, Moonen's presentation calls for a reconsideration of mental imagery, classically considered a purely internal process, as fundamentally action-oriented. Finally, Prof. Slagter will discuss findings from a series of studies showing that action plans tune working memory representations, biasing future performance, demonstrating that working memory does not simply serve perception beyond the immediate, but also serves adaptive future behavior. Together, the presentations in this symposium emphasize that even high-level cognitive processes are deeply rooted in sensorimotor processes and should be understood as such.

This symposium reflects the growing momentum across philosophy, psychology, and neuroscience toward approaches that recognize the central role of the body, action, and the world in shaping cognition. It emphasizes the reciprocal relationship between empirical psychological research and its philosophical foundations, calling upon philosophers and psychologists to work together to advance understanding of how even high level cognitive processes depend on the body's action systems and the affordance structure of the environment.

Julian Kiverstein: *Making sense of multi-functionality in the brain: an action-oriented cognitive ontology*

Much of cognitive neuroscience has been concerned with the mapping of specialised cognitive and affective functions onto regions and circuits of the brain. Brain-imaging techniques such as fMRI are used to map task-specific patterns of brain activity to behaviour and cognitive functions. What does it mean to claim that a brain region has a specialised cognitive, affective or behavioural function? This talk will take up the recent arguments of Luis Pesoa (2023) that the units of function in the brain are not individual areas or regions but networks of regions (cf. Anderson 2014). Pessoa has argued that brain regions dynamically assemble into coalitions and populations that realise complex cognitive-emotional behaviours. Functional relationships between regions fluctuate based on cognitive, emotional and motivational demands. The same regions can belong to different networks at different times and thus be recruited to perform multiple cognitive functions (Anderson 2010, 2014; Klein 2012; Viola 2017; McCaffrey 2023). The function a region performs changes on the timescale of seconds depending on the coalitions and relationships it forms with other regions. That is to say, by the structure of the whole network.

I will consider the implications of Pessoa's argument for the question of cognitive ontology – the psychological categories that guide investigation in cognitive neuroscience and psychology more generally. First, I will argue for a bottom-up approach to cognitive ontology that seeks to derive cognitive categories from neural categories. By studying the variety of functions brain regions can perform, we can answer the cognitive ontology question of which cognitive categories can productively guide research in cognitive neuroscience and psychology. Second, I will argue that the cognitive categories required to make sense of how brain regions can perform a variety of functions relate to action control and planning and thus cut across traditional distinctions between

cognition, affect and motivation (Hurley 2008; Anderson 2014; Pessoa 2023). According to this action-oriented perspective, the nervous system originally evolved for sensorimotor coordination and these action-oriented biases have been preserved in the human brain. For example, basal ganglia-cortical loops bidirectionally connect cortex with sub-cortex allowing cortex and subcortex to work in close coordination. Basal ganglia-cortical (BG-C) loops have been found to be preserved across vertebrates – reptiles, fishes, birds and mammals. BG-C loops are an example of a functionally integrated network that serve an action control function. However, the BG-C loops that form in birds and mammals are more elaborate suggesting that a functionally integrated circuit that originally served an action control function can be recruited to support more elaborate and complex cognitive behaviours. I will argue on this basis for an action-oriented cognitive ontology.

Iris Groen: *Gibson's neural reality: locomotive affordances in the human brain and AI models*

To move, traverse or transport ourselves in the world around us, we can use different actions, such as walking, swimming, or climbing. Ecological psychologist Gibson (1979) famously proposed that affordance recognition, i.e. the process of determining which actions are possible given the current environment, fundamentally shapes visual perception in visual organisms, including humans. The neural basis of this process in the human brain has remained mostly elusive, however. Excitingly, the advent of modern neuroimaging technology allows us to now look ‘inside’ the brain and to identify brain areas and neural computational cascades that may underlie our ability to recognize environmental affordances.

I will present neuroimaging evidence for neural computation of affordances in the context of natural scene perception, as measured with functional magnetic resonance imaging (fMRI) and electro-encephalography (EEG) in humans. We operationalized affordance perception as a simple multiple choice task, in which participants were shown a picture of an unfamiliar, real-world environment (e.g., a mountain trail; a bridge crossing; a bouldering hall) - and asked to indicate which actions they would undertake to move in such a scene. We then measured their brain activity to assess, when confronted with each visual scene, how the brain knows what to do – which brain responses reflect the locomotive actions the environment affords?

Via careful comparison of the measured neural activation patterns with various models of scene information, we show that specific cortical regions and temporal processing stages uniquely reflect locomotive action affordances. We find that multivoxel fMRI responses in regions of visual cortex known to be involved in scene perception represent perceived locomotive affordances, and do so independently from other scene properties such as objects, surface materials, scene category, or global properties, and independent of the task performed in the scanner. Similarly, analysis of millisecond-resolution time-courses of EEG responses evoked by the visual scenes reveal that locomotive action affordance representations emerge within 200 ms of visual processing, with again independent unique contributions at temporally distinct time-points from other properties. These empirical observations uncover novel evidence for neural representations in the human brain that reflect locomotive affordances.

Finally, AI has seen tremendous progress in recent years through machine learning on large corpora of pictures or text. However, when engineering real-world interfacing AI in e.g. robotics, incorporating affordance recognition has been difficult. To understand why, we compared a suite of AI models to the behavioral and brain responses from our human participants. We demonstrate that commonly used models of visual processing in human brains, namely deep neural networks trained on scene understanding tasks, do not adequately capture the unique representations of affordances we observed in humans. We propose that this discrepancy reflects human

embodiment, which allows affordance representations to emerge in the brain, but which current architectures of scene understanding AI models lack.

Together, these findings suggest that human locomotive affordance recognition relies on specialized neural representations different from those used for other visual understanding tasks, and different from common AI models.

Lydia Moonen: *Toward a functional, action-oriented approach to visual mental imagery*

Visual mental imagery is commonly thought to be experienced in the mind's eye, without any overt behavioural responses. Accordingly, cognitive neuroscience on visual mental imagery has predominantly focused on capturing sensory representations, rather than incorporating behavioural paradigms to explore its functional role in guiding our actions. This emphasis stands in stark contrast to how cognitive processes, such as memory and decision-making, are studied, which are typically operationalized through behaviour.

We argue that prioritizing sensory content over behavioural effects in visual mental imagery research has led to visual mental imagery being conceptualized as a passive mental process that operates independently of other cognitive processes, thereby isolating it from the broader cognitive architecture. This focus hinders significant progress in leading debates within visual mental imagery research, which largely concern the relationship between visual mental imagery and other cognitive processes. Specifically, we propose that a functional, action-oriented account of visual mental imagery can help clarify (1) whether mental imagery is similar to perceptual processing and visual working memory, and (2) whether visual mental imagery can be unconscious. Including behavioural paradigms in visual mental imagery research is therefore crucial for providing a functional account of mental imagery, supporting meaningful advances in current debates, and enabling a genuine integration of this mental process within the cognitive neurosciences.

Heleen Slagter: *Visual working memory for action*

Philosophers and psychologists have long proposed that perception is organized around action: that we selectively process sensory features that matter for what we can do next. Empirical research increasingly supports this selection-for-action view for objects currently in sight, though it remains underappreciated in many standard accounts that still consider perception and action independent processes. An open question, however, is whether action planning similarly structures the representations we maintain in mind, in working memory. This is the question my talk addresses. Traditionally, visual working memory has been approached as a temporary storage system for previously seen information. In my talk, I will present convergent evidence from behavioral, eye tracking and EEG studies showing that representations in working memory are tuned by what we plan to do with that information next, emphasizing that visual working memory fundamentally serves future behavior. In a first set of studies, we found that when participants planned actions on an object held in mind, this strengthened their sensory representation, as reflected in enhanced attentional capture on an intermittent visual search task (shown with eye tracking) and a larger Ppc, an ERP component associated with visual saliency (shown with EEG). In a second set of studies, we manipulated whether two oriented bars held in working memory were associated with the same action plan or two different action plans. We found that action similarity affected the extent to which the two bars, when similarly oriented, were reported as more visually distinct. Moreover, a follow-up study showed that consistent object-

action relationships, learned over many trials, amplified these perceptual biases, suggesting that action planning structurally warped "visual" feature space in working memory.

Together, these findings indicate that working memory representations are tuned to the actions we intend to perform. They show that visual working memory does not simply serve perception beyond the immediate, as traditionally assumed, but is geared towards adaptive interaction with the world. More broadly, they highlight the deep interdependence of action and cognition.

Wednesday 1st July 14:30 — Imagination & Hallucination (Grote zolder)

Eva Rafetseder, Julia Wolf and Josef Perner: *Pretending Something Exists Is Easier Than Pretending It Doesn't: What Pretence Reveals About Thinking of Non-Existence*

Pretend play is one of the earliest ways in which humans engage with non-reality. Children readily pretend that an empty cup is full, that a stick is a sword, or that a block is a car. Yet much less is known about how the mind represents the opposite kind of situation: pretending that something which is actually present does not exist.

This talk explores the theoretical significance of a newly observed asymmetry between these two forms of pretence, which we call Positive Pretence (pretending that something exists when it does not) and Negative Pretence (pretending that something does not exist when it does). In pilot data from 2- to 6-year-olds (N = 72), children were significantly more accurate in Positive Pretence than in Negative Pretence, even when controlling for age and general task demands. For example, children found it easier to pretend that an empty container held an object than to pretend that a visibly present object was absent. This asymmetry is robust, novel, and not predicted by existing theories of pretend play. Theoretically, this finding raises a fundamental question: Is pretending about non-existence representationally more demanding than pretending about existence?

To address this, I draw on the Mental Files framework, which models thought as structured around mental "files" that track objects and store information about them. While this framework has been fruitfully applied to reference, perspective-taking, and false belief, it has not yet been systematically extended to non-existence. I explore two broad classes of explanation. On one view, Negative Pretence requires a special representational operation—such as blocking or negating an existing mental file—which introduces additional cognitive demands. On an alternative view, Negative Pretence leaves key aspects of representation (such as location or existence) under-specified, making it vulnerable to interference from reality. Both accounts suggest that pretending that something does not exist is not simply the mirror image of pretending that something does exist. The talk argues that this asymmetry is not a mere performance effect but reveals a deeper structural feature of how the mind represents absence versus non-existence. More broadly, the pilot data motivate a rethinking of pretence, and open up a new empirical route for investigating how humans think about what is not there.

Daniel Kim and Keith Allen: *Anticipation and Illusion*

Naïve Realism is a prominent contemporary Anglophone theory of perception according to which perceptual experience consists in a direct, non-representational relation to aspects of the mind-independent world (Campbell, 2002; Martin, 2002; Brewer, 2011; Soteriou, 2013; Allen, 2016; 2019; Kim, 2022; 2024). While this view is often thought to best capture what a perceptual experience is like from a first-person perspective, critics have long objected that it cannot adequately explain perceptual error, including hallucination and illusion. Much of the recent literature has focused primarily on hallucination and on whether naïve realism can accommodate it via a disjunctivist strategy (Martin, 2004; Fish, 2009; Soteriou, 2016). By contrast, the problem

of illusion has received comparatively little sustained attention. Yet illusions are distinctive: unlike hallucinations, where there is no appropriate object of perception, they are typically understood as genuine cases of perceptual experience in which mind-independent entities are present, even though those objects appear other than they really are. The central challenge for the naïve realist, therefore, is to explain how illusions can be both genuinely world-involving and non-veridical or erroneous (Millar, 2015).

This paper develops and defends a unified account of illusion – the anticipatory view – which integrates core commitments of naïve realism with ideas from the Phenomenological tradition, especially Husserl's notions of horizon and fulfilment (Husserl, 1907/1997; 1913/2012; Merleau-Ponty, 1945/2012; Madary, 2012; Textor, 2018). On this view, perceptual experience is essentially structured by a horizon of implicit anticipations concerning how an object would or might appear under different conditions and from different standpoints, together with the ways in which those anticipations are fulfilled or disappointed. In veridical perception, anticipations are continually fulfilled through ongoing acquaintance with aspects of the mind-independent world, thereby grounding the subject's perceptual confidence and familiarity. Illusion arises when this harmony breaks – at least temporarily – because the object's current perspective-dependent appearance, with which one is directly acquainted, conflicts with prior or ongoing anticipations about its other possible appearances. Some anticipations are thereby annulled as invalid, and the object is experienced as having 'mere appearance' rather than as being as it really is.

The anticipatory view contrasts with two standard naïve realist treatments of illusion. First, some naïve realists adopt a doxastic strategy, holding that illusion occurs when one perceives an object while failing to perceive certain of its properties and subsequently form a false judgment that the object instantiates a property it in fact lacks (Fish, 2009). For example, in viewing a round coin through a distorting lens, one might be said not to perceive its roundness but to experience it as elliptical, and then to judge – erroneously – that the perceived object is elliptical. Second, others appeal to the notion of looks or appearances (as aspects of mind-independent objects), suggesting that the perceived object instantiates a look that is characteristic of a kind to which it does not belong (Brewer, 2011; Genone, 2014). Thus, when a white wall is viewed under red lighting, the wall has a look characteristic of red things. The experience is illusory not because one experiences the wall as red, but because one is acquainted with a look typical of red objects that the white wall happens to instantiate in those conditions.

Millar (2015) argues that both strategies are inadequate. The judgment-based approach fails because it locates error at the level of judgment or belief rather than in perceptual experience itself, thereby stripping illusion of its distinctive 'perceptual' phenomenology. The look-based approach fails because perceiving a look characteristic of F-objects while perceiving a non-F object is not sufficient for experiencing the object as F. A coin viewed from an oblique angle may possess looks typical of ellipses, but that does not entail that it is seen as being elliptical. According to Millar, once this insufficiency is acknowledged, the look-based view collapses into the judgment-based view and inherits its problems.

The anticipatory view avoids these difficulties. Unlike doxastic accounts, it explains the erroneous character of illusion at the level of perceptual experience itself, rather than relegating error to post-perceptual judgment. Unlike standard look-based views, it construes looks or appearances not merely as situation-dependent features of objects or environments, but as perspective-dependent properties embedded within a dynamic structure of anticipation and fulfilment. This provides richer phenomenological grounds to accommodate a wide range of illusory cases, including those in which both environmental and subjective factors contribute to phenomenology. Some illusions may stem primarily from features of the environment that plausibly mislead, while others may depend more heavily on the functioning (or malfunctioning) of the subject's perceptual system.

Moreover, the anticipatory framework promises a unified treatment of related phenomena such as perceptual indeterminacy and amodal completion, all within a broadly naïve realist framework.

Silvana Pani: *Can We Share Mental Imagery?*

Mental imagery plays a central role in many forms of human cognition, yet its place in interpersonal coordination remains unclear. Unlike perception, it is not anchored to publicly accessible stimuli, and unlike propositional attitudes, it lacks clear criteria for when two individuals are in the same mental state. This might lead to the view that while imagery may accompany communication or joint action, it cannot itself be genuinely shared, given its standard characterization as perceptual processing not directly triggered by sensory input (Pearson et al. 2015; Dijkstra et al. 2019; Nanay 2023). This paper challenges that conclusion. It argues that sharing mental imagery is possible without shared imagistic tokens or fully identical imagistic contents, by distinguishing three senses of sharing: by token identity, by content identity, and by structural coordination of cognitive activity.

On the first sense, two subjects would share a mental image only if the same imagistic token occurred in both minds. This is implausible: mental images are instantiated in individual subjects, and there are no obvious criteria for re-identifying a numerically identical imagistic token across minds. Moreover, the high variability of mental imagery, even at the level of conscious report, undermines any such attempt.

On the second sense, sharing would require fully identical representational content. This standard is overly demanding. Imagistic content is often indeterminate along dimensions such as perspective, scale, orientation, and detail, and there is little reason to expect convergence on fully identical content in successful coordination.

These difficulties motivate a third, weaker but functionally relevant sense of sharing. On this sense, subjects share a mental representation insofar as they engage in coordinated imagistic activity structured by the same representational constraints. What is shared is not a mental object or determinate content, but a way of imagining: a format-governed activity regulated by task-sensitive constraints on transformation, elaboration, coherence, and content generation.

Several accounts converge on a constraint-based framework of imagination (Kind 2016; Langland-Hassan 2016; Dorsch 2016; Myers 2021; Gauker 2024). Imagining is regulated by world-preserving, knowledge-based, and format-specific constraints that determine permissible transformations. Coordination is achieved when agents' imagistic activities are jointly answerable to the same constraints, even if their images differ in some dimension, such as vividness or detail (Marks 1973; Zeman 2024). Although mental imagery typically lacks an external and sharable stimulus, it can be aided by public scaffolds such as diagrams, gestures, and shared descriptions. Recent work on shared perception shows that coordination does not require shared experiences, but shared orientation toward what counts as commonly available and relevant (Deroy et al. 2024).

A similar point applies to shared imagery. I contend that imagistic coordination parallels perceptual coordination for joint action and is best explained by a notion of commitment (Bratman 1993). Subjects can be committed to regulating their imagistic activity in light of shared constraints and purposes, as when they discern constellations or plan routes with maps and sketches. In such cases, coordination depends on imagistic operations -- simulating trajectories, comparing spatial relations, adjusting perspectives -- that are not captured by shared belief alone.

Drawing on work on non-propositional motor representations in joint action (Sinigaglia & Butterfill 2022; Mattei 2025), the paper concludes that while imagistic tokens and phenomenal character

remain private, mental imagery is shareable in a functionally relevant sense. What can be shared is a way of imagining: a format-governed cognitive activity that can be sustained across agents and that supports interpersonal coordination even in the absence of shared experiences or publicly accessible stimuli.

Adriana Alcaraz Sánchez: *Imagining oneself in a parallel world: The role of imagination in Dissociative Identity Disorder and Maladaptive Daydreaming*

Dissociation is defined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V; APA, 2013) as a disruption, or lack of integration, of psychological functions affecting memory, attention, perception, affect, behaviour, and identity. A broad consensus in the contemporary literature holds that dissociation is a dimensional phenomenon, spanning from non-pathological experiences commonly found in the general population to severe and disabling dissociative disorders (cf. Soffer-Dudek & Somer, 2022). Thus, dissociative traits are not pathological per se, but may, under certain conditions, lead to significant impairment in daily functioning. Within this context, several authors have recently proposed that certain forms of fantasising might fall within the pathological side of the dissociative spectrum. A prominent example is maladaptive daydreaming (MD), a phenomenon characterised by highly immersive, recurrent, and difficult-to-stop fantasising (Somer et al., 2002). Individuals with MD describe experiences strikingly reminiscent of dissociative psychopathology: intense absorption in internal events, diminished responsiveness to the external environment, and feelings of detachment from reality and from oneself (Soffer-Dudek et al., 2025; Ricci et al., 2025). According to some (Soffer-Dudek & Somer, 2022), the transition from ordinary to maladaptive daydreaming involves a disintegration of consciousness, leading to a “split” or division of the normal stream of consciousness, as exemplified by extreme cases of dissociation such as dissociative identity disorder (DID).

These proposals, which situate well-known dissociative disorders such as DID alongside less well-known disruptive forms of fantasising such as MD, emphasise the role of imaginative involvement in the emergence and maintenance of dissociative psychopathology. In fact, several authors have focused their attention on the trait of dissociative absorption, also known as “absorption and imaginative involvement” in the Dissociative Experiences Scale (DES; Bernstein & Putnam, 1986). This psychological construct captures the tendency to become fully immersed in external and internal stimuli, including daydreaming and fantasising, leading to a state of narrowed attention, automaticity, and reduced awareness of one’s surroundings. High levels of dissociative absorption have been strongly linked to heightened imaginative abilities, including a tendency towards imagination and vivid imagery (cf. Bregman-Hai et al., 2018). However, contrary to other models, which entertain the possibility that dissociative disorders might be merely iatrogenic and emerge as the result of sociocognitive mechanisms (i.e. suggestibility; Spanos, 1994), imaginative involvement is conceived in these accounts as a mechanism that facilitates escapism from distressing experiences (cf. Soffer-Dudek & Somer, 2022). Despite this growing interest in dissociative absorption, the role of imagination itself remains undertheorised within dominant psychological models of dissociation. For instance, what exactly is the type of imaginative involvement that contributes to pathological dissociation? And is dissociation affected by the degree of engagement in imagination, or the type of engagement that takes place?

In this presentation, I aim to address these questions by introducing a novel framework that considers imagination as a key cognitive mechanism in MD and DID. I propose that MD and DID are marked by a prolonged or recurrent inhabitation of the imaginary mode of consciousness—a mode of intentionality in which one simulates a possible experience (Husserl, 1898/1925). In these states, imagination becomes the primary means through which individuals orient

themselves in the world, regulate affect, and sustain coherence and meaning. In this sense, I argue that MD and DID involve an overreliance on imagination as the preferred cognitive tool.

I develop this framework by drawing directly on Merleau-Ponty's work on imagination as an embodied, action-oriented faculty, structurally continuous with perception and grounded in our body schema. For Merleau-Ponty, perception is not a passive reception of stimuli but an active, embodied engagement with a world experienced in terms of possibilities for action (Merleau-Ponty, 2005 [1945]). According to him, imagination relies on the same embodied structures as perception, but is oriented not towards what is actual, but towards what is possible (or conceivable). Imagination thus allows subjects to explore, rehearse, and transform their ways of inhabiting the world.

I show that, when viewed through a Merleau-Pontian perspective, imagination in extreme forms of dissociation is not inherently escapist or pathological. Rather, it is a fundamental cognitive mechanism that enables sense-making, anticipation, and emotion regulation. As such, certain dissociative states can be understood as conditions in which this imaginative mode becomes dominant, overshadowing other modes of world-relation, such as perceptual engagement. Imagination thus functions as a primary navigation tool, rather than a supplementary or compensatory one. On this view, the maladaptivity of MD and DID should not be understood simply as dysfunction or deficit. Rather, these conditions exemplify what may be described as "misplaced success": the imaginative system takes over and becomes overrelied upon to the detriment of other modes of engagement. In this sense, dissociation in these phenomena reflects not the breakdown of imagination, but its dominance as the primary mode of world-orientation.

This phenomenological account sheds new light on MD, DID, and related dissociative states. Under this framework, in both MD and DID there is a mobilisation of imagination as the primary action schema for navigating the world—one relies on the imaginative mode of consciousness to make sense of one's world. For instance, in the case of DID, imagination becomes a cognitive strategy for organising an affectively ambivalent and disrupted sense of self (cf. Maiese, 2016). In MD, imagination becomes a highly practised and habitual skill that is repeatedly relied upon to manage distress, boredom, or unmet needs (cf. Burrell et al., 2025). Yet, only in DID does imagination loses its anchoring in the actual world, whereas this anchoring remains in MD, giving rise to two distinct phenomena.

By reframing dissociation through a Merleau-Pontian account of imagination, this presentation offers a novel conceptual framework that integrates clinical and empirical psychological research on dissociative absorption with phenomenological positions on imagination. It clarifies the role of imagination in dissociative disorders, illuminates the continuity between maladaptive daydreaming and DID, and contributes to ongoing debates about the nature of dissociation as a disturbance of embodied world-relation, rather than merely a trauma-based or socially constructed phenomenon.

Wednesday 1st July 14:30 — Us (Spiegelzaal)

Bart Geurts: *The evolution of joint action: from quorum sensing to prospective coordination*

Joint activities come in all shapes and sizes: bacterial colonies collectively emit light, chimpanzees use gestural signals to initiate and regulate grooming sessions, and humans act together in line dancing, masses, and barbecues. All these joint activities require their participants to solve coordination problems, which are ubiquitous in human society and throughout nature. Some coordination problems are simple and the procedures for dealing with them are fairly

straightforward; others are highly demanding and solved in ways that are, as yet, poorly understood. Bioluminescence is in the first category; barbecuing in the second.

In order to get a handle on this variety, I propose to home in on the correlation devices that enable coordinated action. Correlation devices are circumstances external to the interaction as such that participants conditionalise their actions on, knowingly or unknowingly (Vanderschraaf 1995, Brandenburger and Friedenberg 2008). I briefly discuss two examples, quorum sensing and signalling, followed by a more leisurely discussion of a recent development: commitment making.

Quorum sensing is a comparatively simple case. In bioluminescent bacteria, light emission is coordinated by reference to the density levels of the "signalling molecules" secreted by the bacteria (Verma and Miyashiro 2013). These density levels serve as the bacteria's correlation devices, which enable them to light up in unison when population density reaches a threshold.

Although quorum sensing is often treated as cell-to-cell signalling, the so-called "signalling molecules" secreted by individual bacteria are not signals in the canonical sense. It is the density of signalling molecules that serves as a correlation device. Individual secretions impact density levels, but are not themselves correlation devices.

In a standard signalling system of the kind invented by Lewis (1969), signals are correlation devices. Signalling systems have evolved many times and in many species, including chimpanzees, who use gestures to coordinate joint activities such as grooming, play, and sex. Most of the gesture types used by chimpanzees have multiple functions and therefore depend on the context for their interpretation (Graham and Hobaiter 2025). Therefore, gesture types serve as higher-order correlation devices that help their users to disambiguate token gestures, which then serve as first-order correlation devices.

It is a distinctive feature of joint action in our species is that we use language to coordinate our activities prospectively, and not just in the moment. Early in March, Betty and Barney agree to spend the first week of May in Sicily. In order to achieve their agreement, they engage in a communicative exchange, which is a joint activity in its own right, that results in a shared commitment to a joint activity in May. This shared commitment also entails that Betty and Barney have to coordinate their actions in the intervening period by agreeing on a travel plan, booking flights, etc. [Reference to author's publication]

Prospective coordination is a recent development and it is a sine qua non for modern human culture and society. It began with the practice of treating each other as having commitments: normative relations between promisers and promisees, for example, which are negotiated in communicative exchanges. Commitments are correlation devices that, created by communicating in the present, enable us to coordinate our future activities over prolonged periods of time.

Viewing joint action in terms of the correlation devices involved, the following conclusions may be drawn. At a general level, it is evident that, over time, agents acquired more control over their correlation devices. Although bioluminescent bacteria impact the density levels of signalling molecules, no single bacterium controls that parameter. By contrast, chimpanzees use gestural signals deliberately and strategically and the same goes for humans who use linguistic signals for commitment making. Second, and relatedly, both chimpanzees and humans employ signalling conventions as higher-order correlation devices, for which there is no evidence in bacteria. Third, we use communication to negotiate commitments: correlation devices of unprecedented power that enable us to flexibly coordinate our activities over large stretches of time.

Axel Seemann: *Joint Motor Action and Social Space*

This paper argues in favour of the hypothesis that certain kinds of joint actions are made possible by what I call "social space". The actions in question are interactions in which a shared motor goal is pursued by two or more agents who synchronously coordinate their bodily contributions. Pertinent examples include, amongst many others: rowing a canoe together (Knoblich & Sebanz, 2006), playing a piano duet (Wolf et al., 2018), or the many scenarios that are discussed in the context of collective intentionality, such as going for a walk with someone (Gilbert, 1990). Social space is a framework that enables joint action. The hypothesis is that in joint action, the areas around each participant are peripersonal spaces in which sensorimotor information is integrated so as to make motor coordination possible.

The main argument (and organization of the paper) is as follows. Some social creatures, including people, deploy joint know-how in acting together. On one plausible view, joint skill consists in knowing how to move so as to prepare the ground for one's co-agent's contribution to a joint task (Birch, 2018). This requires predicting the other person's contributing movements as well as one's own. Making such predictions relies, for each participating agent, on the representation of one's partner's as well as one's own future movements. The question then arises how these representations can be integrated into a unified structure that is conducive to skilled interaction. Sinigaglia and Butterfill (2020, 2021) suggest that agents rely on "agent-neutral representations" that enable them to develop non-accidentally matching plan-like hierarchies of motor representations in each agent. Much will depend on how the notion of agent-neutrality is interpreted. I opt for a weak interpretation in terms of "agent symmetry", according to which do the representations of joint motor goals specify the kinds of bodily movements by way of whose execution the joint motor goal is realized and thus the kinds of agents whose movements are apt to realise these goals.

I then ask how the agent-symmetrical representation of joint motor goals is possible and appeal to the concept of peripersonal space in answering the question. In peripersonal space, agents integrate information internal and external to their bodies to represent and perform motor actions. One possibility, suggested by e.g., De Vignemont and Iannetti (2015), then is that joint agents use their bodies to represent others' contributions in their own action space. I argue against this possibility, on the grounds that joint agents effortlessly cooperate on objects towards which they occupy different standpoints and that in such cases predictively representing a partner's contribution to a joint motor action requires cognitively costly mental rotation. Instead, I suggest that agents operate in social space, in which they integrate sensorimotor information at both their own and their co-agent's locations (see Seemann (2019) for an earlier treatment). Two joint agents then each operate with a spatial framework in which their own and their partner's locations serve as origins of action. I briefly discuss some experiments that at least tentatively support this view (e.g., Maister et al., 2015; Teneggi et al., 2013), as well as some pertinent evidence from psycholinguistics (Peeters et al., 2015; Peeters & Özyürek, 2016). The appeal to social space triggers the question how joint agents' contributions are unified into one representation of their joint motor goals, so that these contributions are represented as subserving one complex action rather than as two separate ones. I suggest that this is accomplished via a joint body schema (Soliman et al., 2015). The body schema represents in a sensorimotor format parameters that support action execution and control (De Vignemont et al., 2021). The joint body schema draws on proprioceptive information from the agent's body to model the co-agent's contribution at the place occupied by that agent and in this way integrates both agents' contributions. The hypothesis of social space, in conjunction with appeal to the joint body schema, thus explains how agent-symmetrical representations of joint motor goals are possible.

This outcome is presented as a hypothesis in need of further empirical investigation. The benefits of such an investigation are significant: the hypothesis, if confirmed, can make valuable

contributions to debates about social cognition, joint action, perspective-taking, and agent recognition.

Deborah Marber: Imagination and the motivation of (social and collective) action

Pretence cases have been taken by some (see Velleman 2000, Currie & Ravenscroft 2002, Gendler 2006, Doggett & Egan 2007, Van Leeuwen 2016 and Ichino 2019) to show that some imaginings can motivate action like belief. This is at odds with the motivational view of belief, according to which belief's role in motivating action alongside desire is unique. But a recurring objection, Belief-mediation, contends that imaginings' motivational role in such cases relies on intermediate beliefs (see, O'Brien 2005, Schellenberg 2013). I show that the problem of Belief-mediation is yet to be satisfactorily reckoned with and offer a solution through a revisionary account of motivation, call it the Proustian model of motivation.

According to this dynamic model, motivation always results from a positive metacognitive evaluation of an imagined mode of action which is constantly updated in light of new information. This metacognitive evaluation of a mode of action, call it endorsement, of which means-end belief is a paradigmatic example, can have both an affective and a conceptual dimension. Motivation, I argue, crucially relies on the affective aspects (cf. Proust 2013) of this metacognitive evaluation of action modes (system 1), though there are noteworthy connections between these and conceptual judgements through which we can also evaluate action (system 2).

I show why social and collective action contexts, such as that of pretence, elicit feedback loops that strengthen endorsement of (assent towards) some imagined modes of action in a way that enables them to motivate like belief and suggest that this feedback mechanism explains why pretence cases have appeared especially noteworthy to those wanting to defend the claim that imagining motivates like belief.

Next, I assess the implications of this Proustian model of motivation for metaethical constructivism and fictionalism. In particular, I focus on offering an account of metaethical constructivism that takes imagination seriously as the core state at the heart of our judgements of value. My account, call it Imagination-first constructivism builds on Christine Korsgaard's (2009) idea that the norms of actions derive from the need for self-constitution that should be endorsable by the agent at any point in their lives, and so universalisable but puts the onus on the specific communities in which the valuing self is embedded instead of on Kant's categorical imperative.

I also show how it nevertheless allows for decision-makers to fail to be wholly unified as agents even though deciding from a unified perspective (i.e. how it allows for inner conflict) – Imagination-first constructivism requires only imagining de se a possible world that focuses on the particular features of the action-to-be-performed and so involves a rationality that is bounded (cf. Gigerenzer 2010). Finally, it addresses David Enoch's schmagency objection against Korsgaard (roughly, the objection that there is no reason for me to be trying to be an agent when acting) by highlighting how any intentional behaviour requires us to evaluate an imagined action-mode and the role of imagination, propositional and imagistic, in evaluating the consequences of our behaviour."

Yair Levy: Joint Action Without Joint Intention (or Goal)

See PDF.

Wednesday 1st July 14:30 — Belief (Voorkamer)

Robert Ross: *People who endorse conspiracy theories don't necessarily believe them: Implications for survey research*

Survey research on conspiracy theories has grown markedly over the past several decades. This work has produced estimates of the prevalence of belief in conspiracy theories and of numerous correlates, causes, and consequences of these beliefs. A critical assumption underpinning much of this research is that if someone endorses a conspiracy theory in a survey study they sincerely believe it. However, this assumption is rarely tested.

In this presentation, I will summarise the results of two survey studies of belief in conspiracy theories that directly challenge this sincerity assumption (Ross et al., 2026; Williams et al., 2026). In addition, I will argue that these results raise serious questions about how surveys have been used to study beliefs and attitudes more broadly, particularly in contexts where participants might be tempted to respond insincerely to troll researchers or to signal support for their group.

Paul Engelhardt, Dimitra Lazaridou Chatzigoga and Eugen Fischer: *The role of context on the suppression of belief inferences from appearance verbs: Evidence from eye tracking and individual differences*

This study examines a psycholinguistic explanation of fallacies of equivocation and addresses three larger questions about polysemy processing: (1) How do default comprehension inferences contribute to the processing of polysemes? (2) How strongly do default inferences from polysemous verbs influence comprehension, and do they occur despite conflicting context? (3) What individual differences modulate the impact of context on these inferences? Three experiments combined plausibility ratings with eye tracking or individual difference assessments, to address these questions for polysemous appearance verbs, which play a key role in philosophical debates, but have received little attention in psycholinguistics. We argue that verbal reasoning is grounded in automatic language processing and suggest fallacies of equivocation may be partially due to polysemy processing.

Fallacies of equivocation occur, e.g., when people draw inferences supported by a dominant sense of a polysemous word from premises that use this word in an infrequent sense. We focused on fallacies in historically influential philosophical arguments. Arguments from Illusion use polysemous appearance verbs (Robinson, 1994). In their dominant sense ("The car looks small to Claire"), "look", "appear", and "seem" function as subject-raising verbs and attribute to the patient (Claire) attitudes including beliefs about the agent (Looking at the car, Claire sees it is small and believes it is small) (Brogaard, 2013). These arguments contain familiar situations of non-veridical perception (e.g., distance and perspective), where no one believes that, say, the object is as small as it appears from that distance. The arguments rely on a subordinate "phenomenal" interpretation of the verb, which cancels belief implications (Maund, 1986). We followed up on the suggestion that the arguments rely on contextually inappropriate default belief inferences (Fischer et al., 2021).

Irregular polysemes initially activate an internally structured representation of semantic information that is used to interpret the word in different senses (e.g., Macgregor et al., 2015; Brocher et al., 2018). Where a subset of the activated information is relevant for interpreting a subordinate use, it will be interpreted by retaining relevant information from the activated representation and suppressing irrelevant information (Giora, 2012). Where the dominant sense

is much more salient than subordinate senses, complete suppression of irrelevant information is difficult (Fischer & Sytsma, 2021). This motivates three hypotheses:

H1 Phenomenal uses of appearance verbs are interpreted with the Retention/Suppression Strategy. H2 Phenomenal uses of appearance verbs trigger belief inferences that are supported only by the dominant sense. H3 Belief inferences triggered by phenomenal uses influence comprehension.

Three studies examined H1-H3, and whether H2-H3 hold even when the verb is preceded by disambiguating context that suggests a phenomenal interpretation, namely, by specifying non-veridical viewing conditions. Norming studies identified familiar conditions of veridical perception (where, things look their true size, shape, or colour), non-veridical perception (where things look different), and 'neutral' conditions (where one cannot tell either way). In two eye-tracking experiments, participants read three-sentence stimuli and rated their plausibility. In a within-subjects 2□2 design, we manipulated veridicality in the first sentence and consistency with the belief inference ('small' vs 'large', see example below) in the third sentence. Experiment 1 (N=45) contrasted non-veridical with veridical contexts. Experiment 2 (N=48) contrasted non-veridical with 'neutral' contexts. We measured reading times in five regions.

Table 1. Regions of Interest The car in the valley was far away1. It looked2 small3 to Claire4. She believed it was large5. 1Pre-verbal context 2Source verb 3Source adjective 4Source object 5Conflict adjective

H1 predicts higher rereading times on the source verb, when given a phenomenal vs dominant interpretation. H2 predicts higher rereading times in the inconsistent vs. consistent condition for the source and conflict regions. H3 predicts lower plausibility ratings for items in the inconsistent than consistent conditions.

For H1, we compared rereading times between participants who achieved a phenomenal interpretation (evidenced by giving higher plausibility ratings for inconsistent than consistent items in the non-veridical condition) and the remaining participants. In Experiment 1, too few participants responded that way. In Experiment 2, we found a main effect of group at the source verb: $t=-2.16$, $p<.05$ and the source adjective: $t=-2.24$, $p<.05$), indicating that "phenomenal interpreters" went back and reread the second sentence more. Rereading times also supported H2, with main effects of consistency (INCON □ CON) for both source verb (Exp.1: $t=3.32$, $p=.002$; Exp.2: $t=1.83$, $p=.07$ and for source adjective $t=3.11$, $p=.002$ and source object $t=-2.40$, $p=.02$, and for conflict adjective (Exp.1: $t=2.94$, $p=.005$; Exp.2: $t=2.36$, $p=.02$).

Plausibility ratings showed a context x consistency interaction. In Exp.1, mean ratings were INCON □ CON, for both veridical and non-veridical contexts, as predicted by H3. In Exp.2, this difference was observed only for neutral. Since Exp.1 and 2 used the same non-veridical items, we inferred that the more difficult neutral items acted as reflection prompts promoting deeper processing (Alter et al., 2013). Total reading times supported this suggestion. We further inferred that in this more difficult task setting only the more reflective participants managed to suppress belief inferences.

Experiment 3 (N=99) combined the plausibility rating task from Experiment 2 with individual differences measures (Need for Cognition NCS, Digit Span, Cognitive Reflection CRT, Stroop) to assess H3 and H2. A factor analysis revealed that NCS and Digit span loaded on one factor, whereas CRT and Stroop loaded on another. H2 predicts correlations between participants' factor scores for this "reflectiveness-inhibition" factor and plausibility for non-veridical (negative for consistent, positive for inconsistent). H3 predicts that "unreflective" participants will rate consistent items more plausible than inconsistent items, in both neutral and non-veridical, whereas "reflective" participants will do so in neutral, but not non-veridical. Correlations showed (1) the

pattern predicted by H2 (consistent: -.22*; inconsistent: .27**), and (2) the pattern predicted by H3 when splitting the sample on the "reflectiveness-and-inhibition" factor.

We conclude: (1) Polysemous appearance verbs are processed with the Retention/Suppression strategy. (2) Belief inferences influence comprehension and occur despite disambiguating context that undermines them. (3) This influence can be mitigated by stimuli reflection prompts and by comprehender's reflectiveness.

Nicolas Porot: *The Lower Bounds of Belief*

Recently, much attention has been given to which species possess perceptual capacities (Burge, 2010), consciousness (Birch, Schnell, & Clayton, 2020), or episodic memory (Boyle & Burns Brown, 2025). This paper looks at the question of which species possess propositional attitudes (Bermúdez, 2003; Camp, 2009; Carruthers, 2009), focusing on belief. If it is nearly trivial that some animal species have belief, it is less clear which species lack it, or why. I argue a vast number of species are likely to bear beliefs, but that the lower bound on bearing them is fuzzy. The fuzziness is a result of belief's being a natural kind.

I propose two criteria for identifying belief in other species. One is representational format: animals with belief must mentally represent, and at least some of their representations should have language-like structure. Propositional representations make propositional attitudes. Another is belief's psychofunctional profile: animals with beliefs should form, store, and update those representations in ways similar to human belief. There is reasonable evidence for fulfillment of both criteria in several non-human species, suggesting belief is remarkably widespread; uncertainty remains for species that show only partial fulfillment of them.

As concerns propositional structure, many non-human species show capacities diagnostic of a language of thought (reference redacted for review), including logical inference (Pepperberg, et al., 2018; Engelmann, et al., 202; Dautriche, et al., 2023) and abstract representation. There is even tentative evidence for abstract representation in bumble- and honeybees (Solvi, et al, 2020; Howard, et al., 2019). This suggests a large number of species may be on the "having" side of the criterion. Yet, because the diagnostic features of a language of thought are separable ([redacted for review]), it is possible a species may possess some of the diagnostic features, but not all, failing to clearly fulfill the criterion. There is likely a "fuzzy" lower bound to language-of-thought-possessing species.

As concerns the second criterion, comparative cognition is undergoing a phase shift, with rapidly increasing uses of belief-like language to describe non-human species (Whissel, Abramson, & Barber, 2013). At the same time, theorists are developing increasingly rich models of the generalizations that cover how beliefs are formed, stored, and updated in humans (reference redacted for review; Van Leeuwen, 2023). We are beginning to be able to consider specific predictions about belief's role in non-human minds. To take one example, since cognitive dissonance is a recurrent feature of belief updating in humans (cf. Vaidis, et al, 2024), one might expect it to occur in other species, too. Both effort justification (Harmon-Jones, 2017) and spreading of alternatives (Egan et al. 2007) have been observed in numerous non-human species. Another example is performance on false belief tasks. While controversial as benchmarks of theory of mind (Povinelli & Vonk, 2003; Andrews, 2017), success at such tasks provides excellent evidence for an animal's possessing belief more generally (for example, beliefs about another animal's behavior or perceptual states). This criterion, like the first one, may allow for borderline cases, since some of features of belief in humans are separable frome one another.

This approach to identifying belief in non-human animals an example of what is called chauvinism (Block, 1980). We privilege human belief epistemically, measuring mental states in animal minds

against states we know about from human psychology. This rules out mental states in other species that are implemented differently, or are governed by different psychological laws, than our own—but which we might intuitively categorize as propositional attitudes.

Why prioritize the human mind? One answer to this objection is that belief is a natural kind (Sperber, 1997; Van Leeuwen, 2023; reference redacted). Since our best understanding of this kind comes from the human case, human-like features are a non-question-begging starting point for inquiry. It may be the case that over time, our understanding of the principles of non-human cognition shift our understanding of the natural kind, forcing broader characterizations of its most basic features. Over time, our understanding of the natural kind is likely to shift with the full body of evidence on offer.

Mohit Mukherji, Marjorie Rhodes and Eric Mandelbaum: *Mechanisms of belief formation and rejection in childhood*

Adults find rejecting propositions harder than affirming them—they misremember statements that they are told are false as true more often than confirmed statements as false (Gilbert et al., 1990). This is argued to support an asymmetric, 'Spinozan' view of belief formation, on which propositions are accepted implicitly and automatically, but rejecting them requires voluntary effort. This is in contrast to a 'Cartesian' view of belief formation, on which both acception and rejection take the same amount of effort (Gilbert, 1991; Pion et al., 2025).

The developmental roots of this phenomena are unclear. When learning from testimony, children sometimes seem quite Cartesian, rationally evaluating who the experts are, and choosing to believe their testimony over a novice's (Shafto et al., 2012). Other times, however, children seem to have a strong bias to believe and act on others' testimony, even when this testimony contradicts their beliefs (e.g., Jaswal et al., 2010). Importantly, poorer EF predicts whether children are likely to believe conflicting testimony—consistent with the idea that rejecting propositions requires effortful control (Jaswal et al., 2014).

To more directly test these different mechanisms of belief formation in childhood, we presented 3- to 8-year-old children (N=117) with a memory task based on Gilbert et al. (1990). Children were introduced to two characters, a novice and an expert, who were learning about ‘modis’—a new kind of animal. In the “learning” phase, the novice made statements about modis (e.g., “Modis are friendly”), which the expert either confirmed or denied (“Yes, that’s right!”, or “No, that’s wrong!”). Children heard 12 such statements in the learning phase, half of which were denied. After they heard these 12 statements, children were asked 12 questions about the statements they heard earlier (e.g., “Are modis friendly?”). Each of these was a forced-choice yes/no question. Two months later, children repeated the memory test from Session 1 (without any additional reminders or learning phase), and completed a measure of EF (Lagattuta et al., 2011).

We analyzed our data using mixed-effects binomial regression models predicting how likely children were to remember each statement. Children’s EF predicted how accurately they remembered denied statements—but not affirmed ones—in the test phase of session 1, even after controlling for age (interaction between denial and EF, $\beta = 4.59$, $p < .001$). Children who remembered statements correctly during the first session, but misremembered them in the second session were more likely to make these errors for denied statements than confirmed ones (.564, $p = .02$). In sum, children with poorer EF struggle especially at processing and remembering rejected statements correctly. This is compatible with a Spinozan theory of belief, on which accepting propositions is automatic but rejection is effortful.

However, these results were also compatible with an alternative account on which children with lower EF are more likely to answer "yes" as a heuristic to yes/no questions. To rule out this

alternative explanation, we are conducting a follow-up study with another sample of 3- to 8-year-old children. Participants will follow the same general procedure as our previous study, except for two changes. First, half of our participants will see the expert respond to the novice's statement (e.g., "Modis are friendly") without using linguistic negation, by asserting an alternative, contradictory statement ("Modis are mean"). The other participants will hear statements affirmed or negated just as in the first study ("No, that's wrong!"). Second, to rule out the possibility that the "yes" bias uniquely explains why children struggle on denied statements, we will ask children to endorse of two alternative statements on test trials ("Do modis have prickly fur, or do they have smooth fur?"). We are currently collecting data (N = 111; target N = 181).

Preliminary results suggest that when hearing the expert assert a statement that contradicts the novice's statement (instead of simply negating the novice's statement), children with poor EF do not struggle to endorse the correct property for denied statements. This replicates previous work showing that children accept experts' testimony over novices'. But when the speaker rejects a novice's statement by simply negating it, as in Study 1, children with lower EF struggle to remember denied statements more than confirmed ones. This replicates and validates our finding from Study 1 and suggests that a simple "yes" bias cannot explain why children struggle to remember denied statements as false.

Together, our findings are consistent with a Spinozan model on which rejection is effortful, particularly in the absence of alternative propositions.

Wednesday 1st July 14:30 — The Self (Bovenkamer)

Andrea Bertazzoli: *What is the Enactive Account of the Human Self? A Way out of the Body–Social Problem in Cognitive Science*

In the philosophy of cognitive science, the body–social problem is the question, raised by Miriam Kyselo and Ezequiel di Paolo, of how bodily and social dimensions jointly figure in the individuation of the human self as a whole. On the one hand, to overcome the limits of the isolating individualism of much modern Western thought, which conceives of human selves (and minds) as fundamentally separate, self-contained entities, we need to find a way to understand ourselves that recognizes the fundamental importance of our connectedness with others. On the other hand, we cannot however escape the recognition that our identity is grounded in our concrete embodiment in a distinct living body. If we are inescapably social beings, yet undeniably embodied, how can these dimensions be integrated in a coherent account of what we are?

For Miriam Kyselo and Michelle Maiese, the enactive approach to cognitive science, an influential framework that emphasizes both the openness and the fundamental embodiment of living/minded beings, is the best suited to provide a satisfying answer to the body-social problem. They offer two contrasting “enactive accounts of the human self.” Kyselo (2014, 2016, 2020) argues that the human self is primarily a social existence: a self-sustaining network of social relations mediated, however, by the living body. Maiese (2016, 2018, 2022), by contrast, defends the primacy of embodiment. For her, the human self is an embodied organism, normatively shaped, however, by its social environment. Their disagreement revolves around one crucial point: for Kyselo, social relationships are constitutive of our identity; for Maiese, this is not the case and the individual is only socially embedded/shaped, not constituted. Both Kyselo and Maiese appear to pursue something like a singular enactive answer to the question “What am I?”, a unified enactive account of the human self as a whole.

I argue, however, that any such search for a single, fixed “enactive” account of what we are is in tension with the radically constructivist stance toward knowledge that follows from taking seriously the enactive philosophy of Varela, Thompson, and Rosch (1991). At the core of enactivism is the

view of minded/living systems ("selves") as autonomous self-organising systems that bring about and strive to maintain their own precarious identity through their sense-making activity (which includes cognitive and affective processes), that is, through their active bodily engagement with their world. Any form of sense-making, including any production of scientific knowledge, is thus conceived not as an attempt to access and represent a mind-independent world but always as originating from the normative perspective of an autonomous system concerned with its own maintenance. Knower and known are thus revealed to be indissoluble: to know is to have viable ways of acting and thinking within one's world that are functional to achieve one's goals. This radical constructivist stance on knowledge implies then that any scientific account, and thus also any account of the self, is never a static picture of the world assumed to stand on its own, but always only a model constructed by an observer as a tool to guide their thinking in a specific domain of discourse and inquiry.

Once the radically constructivist implications of enactivism are taken seriously – I argue – the body-social problem dissolves and the clash between Kyselo and Maiese on whether the self is socially embedded or constituted reveals itself as ultimately unproductive. The body-social problem arises only if we assume that there must be a single, coherent account of the self as a whole. What Kyselo and Maiese want to model is our "I", but what we call our "I" in ordinary language is not something suited to be identified and discussed at only one level of description and inquiry. It encompasses a huge range of extremely different phenomena – a minimal self-awareness, a sense of one's body as one's own, a sense of one's social identity, and so on – and it arises as a unity, as a single "I", only in the space of language. When treated as competing global accounts of the human self "as a whole," the "enactive" accounts of the human self offered by Kyselo and Maiese risk therefore fixing identity at a single level of description, thereby losing explanatory power by privileging certain dimensions – social or bodily – over certain others.

I conclude by defending the groundless, multilevel view of selfhood offered instead by Francisco Varela. On this view, consistent with enactivism's radical constructivist foundations, identity can be meaningfully attributed at multiple, overlapping scales – biological, affective, neural, interpersonal, political, etc. – without any one of these exhausting what we are. Indeed, identity at one level does not negate identity at another; rather, different explanatory contexts and needs legitimately call for the individuation of different systems as relevant. Bodily and social dimensions are thus not seen here as rival sources of individuation, and the two accounts of the self of Kyselo and Maiese can thus be seen not as conflicting but perfectly compatible, both viable ways of understanding ourselves in different contexts. Abandoning the search for a fixed, unified account of the human self thus ultimately allows enactivism to better accommodate the richness, variability, and plurality of our experiences of ourselves.

Mathijs Geurts: *A New Look at Perspective in Self-talk*

Self-talk, or inner speech, has been discussed extensively by psychologists (McCarthy-Jones and Fernyhough, 2011), philosophers (Martínez-Manrique and Vicente, 2015) and more recently by linguists (Wiltschko, 2025). Literature on the topic commonly describes self-talk as a kind of 'inner dialogue', based on the idea that we can direct utterances at ourselves in the same way we can direct utterances at others (Vygotsky, 1986).

In this talk, I examine the strength of this idea by looking at the function of self-addressed utterances: utterances that refer to the speaker through the use of names and pronouns. The functions of self-talk have been extensively studied, and it has been shown to have positive effects in several areas including sports performance, public speaking and problem solving (Hatzigeorgiadis et al., 2011; Kross et al., 2014; Kim et al., 2021). One striking finding from this literature is that self-talk in the second or third person often has a more beneficial effect on

performance compared to first-person self-talk (Hardy et al., 2019; Furman et al., 2020). An example of this would be saying "You can do it" to yourself instead of "I can do it" to motivate yourself.

I argue that the differences between these different forms of self-talk are best studied within Erving Goffman's participation framework, which describes the different roles that participants can occupy within a speech encounter (Goffman, 1981; Levinson, 1988). The framework distinguishes the different roles participants can occupy in a speech encounter, such as speaker, addressee, or overhearer, and allows us to analyse how self-talk can simulate different configurations of these roles. The idea is that second- and third-personal reference has a special role in evoking encounters with addressees, participants and overhearers.

Using this general framework, I discuss existing explanations of non-first-person self-talk. The dominant explanation of these effects appeals to the concept of self-distancing: non-first-person self-talk promotes a more objective self-perspective (Kross et al., 2014).The self-distancing explanation has recently come under criticism from Rivadulla-Durró (2026), who offers her own 'prima facie' account of self-talk as an alternative. This is the idea that second- and third-person self-talk is implicitly associated with being addressed by others, creating the impression that someone else is making a claim about us. As evidence for her claim, she points out that this view is in line with research on mental imagery processing that shows similar associative effects (Holmes & Matthews, 2010). Both accounts thus explain the performance benefits of self-address and self-reference in terms of a perspective shift, and both link grammatical person to interactional roles within a participation framework, differing only in how they assign those roles. However, in treating second and third person self-talk as instances of the same underlying mechanism, neither account predicts the structural asymmetry that a participation framework reveals: second and third persons occupy categorically distinct roles in conversation

On my account, self-directed utterances that contain addressing devices (e.g. imperatives, vocatives, second person pronouns) or third person reference evoke speech situations with multiple participants, this is shared by second and third person self-talk. The difference is that second personal address is used in order to directly motivate action and that third personal reference is used for self-evaluation. This distinction maps onto different participation configurations: second-person address constructs an addressee role that directly implicates the self as a target of directive speech, whereas third-person reference positions the self as a non-addressed participant, creating the evaluative distance characteristic of self-distancing effects. The current proposal echoes recent work on 'altercentric cognition', showing strong low-level influences on cognition of the present perspective of others (Kampis & Southgate, 2020). Similarly, private speech is both triggered by the presence of others (McGonigle-Chalmers et al., 2014), and can be used to simulate the presence of others. Addressed self-talk may therefore function as a form of simulated social interaction, recruiting the same cognitive mechanisms activated by genuine interpersonal exchange. This suggests that private speech could facilitate both direct and indirect social influence, fostering social learning and later helping to internalise social norms.

Francesco Fanti Rovetta: *The Stabilization of the Past: Feedback Loops in Self-Memory Dynamics*

In this paper, I argue that the relation between self-related information (the self-model) and episodic memories (EM) involves positive and negative feedback loops leading to the stabilization of mnemonic tendencies also known as self-related memory biases. The central claim is that self-related memory biases, such as the self-enhancing bias or the consistency bias (Schacter et al., 2023), are the result of control processes over memory recall subordinated to a hierarchy of

values. In arguing for this claim, I draw from the self-memory system approach (Conway & Pleydell-Pearce, 2000; Conway et al., 2019) and perceptual control theory (Mansell & Marken, 2015). The self-memory system (SMS) is a framework for the integration of episodic information and autobiographical/self-related knowledge resulting in memory retrieval. In line with generative approaches, the framework distinguishes between stored, long-term representations and contextually activated information, and treats both as crucial components (Conway et al., 2019) of memory retrieval.

According to the SMS, single memories are constructed by integrating episodic information about specific events, autobiographical knowledge, and conceptual knowledge about the self, via recurrent and iterative patterns of activation of information until search criteria are met (Conway and Pleydell-Pearce, 2000). Perceptual control theory (PCT) posits that behavior is a function of the attempt to maintain a desired state (Mansell & Marken, 2015). The currently experienced state is continuously compared to an internal goal specification (or reference signal); any discrepancy generates an error signal that drives behaviors to fix the perception. The desired state is influenced by a hierarchy of values, goals, and norms. For example, the value of social safety determines the evaluation of the currently experienced situation with reference to a desired level of perceived safety, and an eventual mismatch leads to taking actions to diminish the mismatch (Gucciardi et al., 2026).

In order to describe the reciprocal relations between self-model and episodic memory, it is necessary to distinguish between 1) the modulation exerted by the self-model on EM at retrieval, considered unidirectionally, 2) the iterative, bidirectional self-memory dynamics and, lastly, 3) the diachronic stabilization of self-memory dynamics resulting in self-related memory biases. With regard to 1), the modulation exerted by the self-model on EM is a process in which information from the self-model is integrated with other information and the memory trace, resulting in the retrieval of a specific memory. From the perspective of perceptual control theory, the influence of the self-model on EM is the result of negative feedback loops that prevent and reduce the mismatch between the actual experience (the recall of a specific memory and related affect) and the desired state, in light of the values of maintaining a coherent/positive/etc. self-model. For example, retrieving a memory that contradicts core beliefs about the self contravenes the value of a coherent self and can cause deeply negative affect, resulting in a great mismatch between current and desired experience, consequently, control processes are involved in avoiding retrieving or modifying such a memory. Conversely, retrieved memories can influence the self-model. For instance, frequent recollection of EMs representing professional successes might result in self-attribution of the trait of being successful encoded in the self-model.

Taken together, the reciprocal influence between self-model and EM gives rise to self-memory dynamics (2). Self-memory dynamics involve iterative feedback loops, such that the self-model at a point in time, S-Mt, shapes EMt, and EMt feeds back into S-Mt+1. In relation to 3), over time and without counterbalancing constraints, self-memory dynamics converge towards inflexible, recurrent, and self-validating patterns. Correspondence, i.e., the tendency to construct memories that are veridical or accurate with respect to factual events they represent (Dings et al., 2023), is one such constraint. Several biases have been indicated in the literature that can be modeled as the result of positive feedback loops in self-memory dynamics. The most investigated is the self-enhancement (or self-serving) bias (Demiray & Janssen, 2015; Schacter et al., 2023), which is the tendency to recall memories that reflect positively on the self. In such a case, a generally positive self-model (S-Mt) contributes to the construction of a positive memory (EMt) which feeds back into the self-model (S-Mt+1) and further enhances it. Analogous processes can be assumed for the negative memory bias observed in depression (Marchetti et al., 2018), for the self-

coherence or consistency bias (Conway, 2005), and others. The stabilization of and interaction between self-related memory biases can be illuminated by PCT.

In particular, the elusive relations between memory biases can be elucidated if understood as resulting from a hierarchy of values affecting control processes, as posited by PCT. For instance, whereas in the general population the bias for self-coherence is adaptive and functional in preserving a positive self-model, in depression, self-coherence supports the maintenance of a negative view of the self, despite the apparent disadvantages of doing so. Although this may seem counterintuitive, it can be readily explained by positing that a coherent identity is prioritized (i.e., it is higher in the hierarchy of values) over a positive self-model. Indeed, evidence suggests that depressed individuals adopt behaviors that reinforce depressive symptoms (e.g., seeking negative feedback about the self from others or assuming self-defeating attitudes in social contexts) in order to validate their negative self-model (Hart et al., 2021). To summarize, I argue that PCT can provide fruitful insights for a systematic analysis of self-memory dynamics and the formation and stabilization of self-related memory biases.

Nicolas Goupil and Dora Kampis: *Episodic recollection before and after self-recognition*

Adult episodic memory involves consciously re-experiencing oneself in contextually rich reconstructions of past events. Developmental research still debates when young children (Newcombe et al., 2014) or infants (Behm et al., 2025) start forming contextually specific memories, binding “what happened” to “where and when it happened”, but there is a consensus that episodic-like memories emerge years before children show reliable self-conscious recollection around the age of four. Yet experiments such as the mirror self-recognition task (Rochat, 2003) suggest that a conceptual self-awareness emerges between 18 to 24 months. Mirror self-recognition has been found to predict memory for a location (Howe et al., 2003), and together with maternal reminiscing style, later episodic memory (Harley & Reese, 1999). One proposed mechanism is that the self-concept may serve as an anchor around which memories can be organized – which later in development also facilitates conscious re-experiencing of episodes (Ross et al., 2025). We hypothesized that although conscious episodic recollection may not emerge before much later, the emergence of conceptual self-awareness might already help early binding mechanisms to form precocious episodic-like memories.

Infants (n=130) participated in the study at 12 and 18 months. At 18 months, self-awareness was measured with the mirror self-recognition task, where around 18 months 50% of infants show self-awareness by touching a mark on their face when looking in the mirror (Amsterdam, 1972). Next, episodic-like binding was measured in the ‘two-room task’ (Newcombe et al., 2014) where children learn the location of two different toys, in two different boxes, in two different rooms. Successful binding is reflected in searching in the box that contained the toy in that room. However, because both rooms contain the same boxes, children sometimes search in the box that contained a toy in the other room; a binding error especially frequent at 18 months. Episodic-like binding was also measured in screen-based eye-tracking tasks at 12 and 18 months, measuring looking time and saccades in response to faces matching (vs. non-matching) scenes they were previously shown with (Richmond & Nelson, 2009), or items matching (vs. non-matching) items they were previously paired with (Johnson et al., 2020) in single exposures. Since some eye-tracking experiments found episodic-like binding already in young infants, we conducted them longitudinally to evaluate their development, and predictivity of binding in the two-room task. Out of N=120 at the 18 months visit N=65 recognized themselves in the mirror, N=52 did not.

One first analysis replicated episodic-like binding in both eye-tracking and interactive paradigms. Participants of the two-room task (N=108) searched the correct box significantly more than

chance ($p < .001$). Participants also showed novelty responses in eye-tracking experiments, with no significant difference between 12 and 18 months. Person-context binding showed in both higher looking times ($p = .043$) and saccades ($p = .034$) to matching faces. On the contrary, item-item binding showed in both higher looking times ($p < .001$) and saccades ($p = .016$) to non-matching items. Altogether, results replicate behavioural signatures of episodic-like binding at the beginning of the second year of life.

Ongoing analyses investigate relations between tasks. A path analysis modelled the development of episodic-like binding in eye-tracking tasks, its relation to binding in the two-room task, and the relation of these to self-recognition. Binding novelty responses, in eye-tracking tasks, showed negative predictions from 12 to 18 months: item-item binding switched from familiarity to novelty preference, suggesting increasing processing efficiency; item-context binding switched from familiarity to novelty preference, which could reflect the opposite, possibly due to processus refinement. Binding in screen-based tasks did not strongly associate with search in the more naturalistic and declarative two-room task, which could reflect the very different task demands, or possibly different memory processes. Most notably, while correct retrieval rates in the two-room task did not relate to mirror self-recognition, error patterns did, with significantly more recognizers committing errors in context-binding (search in the box that contained a toy in the other room), and non-recognizers instead committing random errors (search in box that never contained toys). A pattern that suggests conceptual self-awareness may only indirectly benefit episodic-like memory, either anchoring contextual information only partially (the box but not the room), or generically (across rooms); that is, a still developing capacity for reliable episodic memories.

In summary, episodic-like binding capacities seem to develop before explicit verbal recollection can measure it. Before the offset of infantile amnesia, preverbal infants may already form a wealth of episodic-like, but not yet adult-like, memories. An open question is whether and how such memory may relate to the emergence of autonoetic consciousness, the ability to re-experience oneself throughout contextually rich past events, and to access it explicitly.

Wednesday 1st July 14:30 — Understanding One Another (Kleine zolder)

Sophie Keeling: *Standpoint know-how*

According to the influential standpoint epistemology programme, certain social positions give rise to standpoints that provide their bearers with epistemic advantages. This is to say that standpoints put groups in a position to acquire certain pieces of knowledge, and to talk of knowledge as situated. While not interrogated, the standard use of ‘knowledge’ in this literature is largely taken to be a propositional in nature, that is, the knowledge that such-and-such is the case. For example, it might take the form of knowing that oppression takes a certain form. I will argue that in addition to know-that, groups can possess what I call standpoint know-how.

§1 Introduces standpoint epistemology. Theorists have argued that social positions engender certain standpoints that puts their bearers in a better position to gain certain forms of knowledge. Proponents include Collins (1986), Harding (1992), Hartsock (1998), Fricker (1999) Wylie (2003) Rolin (2009) Pohlhaus (2012), Toole (2020). For example, if one wants to know about the climate of sexual harassment and microaggressions in a workplace, they would do well to ask women and not men. That is to say, standpoints come with an epistemic advantage, such that oppressed groups are better at knowing certain things than those with privilege. Importantly, belonging to a social group isn’t enough to have this advantage. Standpoints aren’t automatic. Rather, standpoints are achievements that must be earned and must be developed as a group. This

process is referred to as ‘consciousness raising’. Standardly, the assumption seems to be that this situated knowledge would be a form of propositional knowledge.

§2 argues that we can better understand many instances of standpoint knowledge instead as know-how, not know-that. Philosophers such as Gilbert Ryle, Jason Stanley and Carlotta Pavese have famously observed that in addition to propositional knowledge, we can also possess know-how. For example, I know how to ride a bike, cook a cake, and recognise different tree species. There is a further debate about whether know-how ultimately reduces to know-that, but I can remain neutral on this.

I will argue that in addition to standpoint knowledge understood as know-that, groups can also acquire standpoint know-how through consciousness raising. One form this can take is as recognitional capacities. Just as skilled doctors can recognise different kinds of illness, the epistemic advantages afforded by certain standpoints consist in the ability to recognise instances of oppression and to recognise oppressions as consisting of a certain kind. For example, it can consist in the ability to recognise microaggressions as microaggressions that would be imperceptible to others, recognising instances of sexual harassment and bad sex, and so on. It is through consciousness raising that groups develop these skills. I will also examine other kinds of socially-situated competences, such as code-switching and emotional labour. These skills are developed by marginalised groups, and again, requires consciousness raising to perform them under these descriptions. That is, performing these skills under the mode of description are achievements of consciousness raising, not just skill acquisition.

§3 then outlines several upshots:

Objectivity

This account allows us to say how standpoint knowledge is situated but also objective. There is a fact about whether a given bird is a blue tit, and similarly, there is a fact about whether something counts as an instance of oppression. One just needs to train and to occupy the relevant ecological niche. Hermeneutical injustice and practical agency My account helps us to locate a new form of hermeneutical injustice. Following Fricker (2007) hermeneutical injustice is the idea that subjects can be marginalised in lacking the conceptual resources to make sense of their situation. I will argue that there are certain skills that subjects can come to perform under a particular mode of description through consciousness-raising. But there will have been a time before the appropriate concepts were developed, such as the concept of ‘code-switching’. As such, hermeneutical injustice also limits the act types that individuals can perform under the relevant modes of description.

Inarticulacy

Appealing to the role of know-how in situated knowledge explains why bearers of standpoint advantage many not always be able to explain every point they make. We can imagine someone noting an instance of harassment or microaggression but being flumouxed at the point of explaining why it counts as such. In this case, we shouldn’t automatically discount them as being uncredible. Experts can’t always explain their skills. The expert doctor can’t always explain why exactly she thinks that the tumour is benign, for example.

Socially-situated competences as expert skill

This presents a new domain of study for philosophers of expertise and skill. A lot of attention has been paid to music, dance, and sports, but I propose that oppressed groups often also develop

high-level expertises. There will be interesting similarities but also differences in the development of these skills, since standpoint know-how isn't the subject of training in the same way.

Amrisha Vaish, Qiao Chai, Parvathy Viswanath, Camile Bernard, Upali Chakraborty and Aneesh Kumar: *The Development of Children's Understanding of Gratitude in Two Cultures*

Gratitude is a positive emotion that arises from the *perception that one has benefited through the kindness or* good intentions of another (McCullough et al., 2001). It serves critical functions in establishing, sustaining, and enhancing cooperation within interpersonal relationships and the broader society (McCullough et al., 2008). However, cultural frameworks influence how gratitude is understood and expressed, and little is known about how these differences emerge during childhood, especially outside WEIRD populations.

The present research examines gratitude development in children from two contrasting cultural contexts: the U.S. and India. At least two key cultural differences between the U.S. and India are relevant to this work. First, Indians tend to view interpersonal relationships and responsibilities as obligations, whereas WEIRD cultures place greater emphasis on their voluntary nature and are thus more inclined to evaluate partners' motivations to decide whether to remain in the relationship (Barrett et al., 2016; Miller & Luthar, 1989). Consequently, the helper's motives may carry more weight for U.S. Americans than for Indians, leading to differential gratitude expectations for voluntary versus non-voluntary benefits in the U.S. more than in India. Second, Indian culture promotes respect for social hierarchy, and proper hierarchical relations are maintained through codified views of reciprocity (Trommsdorff et al., 2007). Thus, gratitude and reciprocity in Indian society are more likely between individuals of similar status than between those of unequal status, whereas status may not impact gratitude and reciprocity in the U.S.

To test these hypotheses, we recruited children aged 5-10 years from middle-class families from across the U.S. (N = 270; total planned N = 300) and Bengaluru, India (N = 276; total planned N = 300). In a 2x2 within-subjects design, participants viewed vignettes depicting a protagonist helping a child (see Figure 1). We manipulated the helper's motivation (voluntary/non-voluntary) and the helper's social status (adult/child). Children were asked whether the recipient would feel grateful toward the helper. The hypotheses and analyses were pre-registered on OSF. Preliminary results align with our predictions.

Children's gratitude expectations revealed a 3-way interaction between Motive x Age x Culture ($ps < .0001$; Figure 2). Specifically, with age, U.S. children were less likely to expect that the recipient would feel grateful for non-voluntary help ($ps < .0001$), but their gratitude expectations for voluntary help remained consistently high across ages. There was a similar pattern observed among Indian children, but it was significantly weaker than among U.S. children (as seen in the 3-way interaction). Thus, as predicted, with age, U.S. children's gratitude expectations showed greater sensitivity to helper motives than those of Indian children. Children's gratitude expectations also revealed a 3-way interaction between Status x Age x Culture ($p = .01$; Figure 3). Specifically, U.S. children's gratitude judgments did not vary based on adult vs. peer helpers; rather, there was only a main effect of age ($p < .0001$) such that older children generally expected lower gratitude than younger children. However, for the Indian sample, we found the expected Status X Age interaction ($p = .014$). With age, Indian children expected the recipient to feel less grateful for receiving help from an adult than a peer. Thus, as predicted, with age, Indian (but not U.S.) children's gratitude expectations showed increasing sensitivity to helper status.

These findings indicate that cultural values shape children's emerging understanding of gratitude. In particular, features that current psychological and philosophical accounts consider central to gratitude, such as the benefactor's motives, may not be universally central to gratitude. At the

same time, features that are not included in current accounts, such as the relative status of helper and recipient, may need to be considered. More generally, gratitude – and likely other social and moral emotions – must be reexamined through a culturally informed lens.

Daniel James, Steffen Koch, Alex Wiegmann, Leda Berio and Kurt Erbach: *Because I'm Not a NAZI? Sociolinguistic Variation and Cross-Linguistic Race Talk*

Allegedly, Charles W. Mills once asked Jürgen Habermas: "Why don't you work on race?" Habermas, frowning, replied: "Because I'm not a NAZI?" This exemplifies an interlinguistic misunderstanding that reflects different attitudes toward race in the United States and Rasse in Germany. Mills's question assumes that a critical theorist has good reason to engage with race. Habermas's response suggests the opposite: only a Nazi would study Rasse. We explore the hypothesis that this and similar exchanges reflect sociolinguistic variation rather than a semantic difference between "race" and "Rasse". What unsettles many German speakers is not what "Rasse" means but who typically uses it and which perspectives it is heard expressing. In this respect, the word behaves in ways familiar from discussions of slurs. Examining this empirically, we conducted several experiments with German and U.S. American participants. Our theoretical framework builds on work on slurs and sociolinguistic variation treating lexical choice as evidence of social positioning rather than as a mere vehicle for descriptive content. Nunberg (2016) argues that slurs – and, we suggest, other socially charged terms like Rasse – convey information about their users, signalling group affiliation, ideological alignment, and complicity with particular worldviews. Generalising this insight, Nowak (forthcoming) emphasises the role of metadata: socially encoded knowledge about a term's histories of use, typical users, and associated perspectives, which shapes how utterances are interpreted and speakers are evaluated. This perspective aligns with sociolinguistic accounts that treat linguistic variation as a pragmatic resource for managing social meaning (Burnett 2019). On such views, word choice positions speakers in ways that invite particular evaluations of stance or allegiance, independently of propositional content. Camp (2013) and Bolinger (2015) show how lexical choices can foist perspectives or agendas onto audiences regardless of speakers' explicit intentions, while Davis and McCready (2020) argue that slurs invoke a complex of sociohistorical facts and attitudes even when used by speakers who do not endorse them.

Taken together, these accounts support a view on which projective or expressive effects are not exhausted by truth-conditional meaning, but can arise from a term's social provenance and conditions of uptake. Seen through this lens, the cross-cultural contrast becomes intelligible: for many German speakers, Rasse is saturated with Nazi provenance and readily heard as affiliating its users with a racialist perspective, whereas in the United States race is closely tied to civil-rights discourse, social science, and law, and is often heard as analytically – or even solidaristically – motivated. These divergent patterns of uptake thus reflect not a semantic difference between the terms, but distinct sociolinguistic ecologies that structure how lexical choices are interpreted and evaluated. In one experiment, German and U.S.-American participants read short vignettes in which a speaker used either race or Rasse to ask a question that could plausibly be interpreted as racist or as anti-racist/social-scientific. Participants rated the speaker on 7-point scales (racist … anti-racist; conservative … liberal). The results revealed a clear intergroup difference: German participants tended to rate the speaker as racist and right-wing, whereas U.S.-American participants did not. Although suggestive, these findings do not by themselves distinguish between a semantic and a metadata-based explanation. On a semantic explanation, Rasse carries racialist truth-conditional content (Ludwig 2018; cf. Appiah 1990): its correct application presupposes that humanity is divided into biologically grounded races whose members share inherited traits extending beyond surface morphology and bearing on normatively salient capacities or dispositions. Because even questions carry existential presuppositions, the use of

Rasse would thereby commit the speaker to the existence of such races, which may suffice to license the judgement “racist”.

A natural diagnostic for our aim is the P-family (projection) test (Chierchia & McConnell-Ginet 2000), identifying content that survives negation, questioning, or hypothetical embedding. However, both existential presuppositions associated with racialist semantic content and socially encoded metadata tied to a term’s provenance can give rise to projection effects. As Tonhauser et al. (2013) emphasise, the P-family test alone cannot determine what kind of not-at-issue content is responsible. At this stage, our aim is therefore not to classify projective content exhaustively, but to test whether the observed speaker evaluations are better explained by responsibility for lexical choice under conditions of sociolinguistic risk than by truth-conditional semantic commitments alone. Importantly, such effects need not be taken to show that Rasse encodes expressive content in the sense of Potts (2007); they can arise from socially salient provenance and asymmetric responsibility for articulation, even where lexical meaning itself remains neutral. We thus examine embeddings of Rasse in direct and indirect speech (Cepollaro et al. 2024). Neither construction commits the reporting speaker to the existence of races: if A reports that B said that C and D belong to different races, or quotes B as saying so, A does not thereby endorse a racial ontology.

However, speakers using indirect speech bear responsibility for their own lexical choices. Because indirect reports permit alternative formulations—paraphrase, scare quotes, or explicit distancing—failure to avoid or mark a socially charged term may be taken to signal affiliation with its provenance. By contrast, direct quotation reproduces another’s words rather than selecting one’s own.

We therefore test the interaction between word choice (Rasse vs. Arschloch) and report type (direct vs. indirect). We predict (i) a significant difference in racism ratings between indirect and direct reports featuring Rasse, reflecting the projection of provenance-based metadata under asymmetric responsibility for lexical choice, and (ii) a substantially weaker indirect/direct difference for Arschloch—a hybrid pejorative whose negativity is lexically encoded—consistent with its evaluative force being less dependent on social provenance (Milić 2018). If these predictions are confirmed, the racism ratings observed in our first study can be shown to arise, at least in part, from metadata associated with Rasse, independently of any racialist semantic content it might encode.

Together, the experiments illuminate what was at issue in the Mills–Habermas exchange and help explain why race discourse diverges across linguistic and national contexts: German unease over Rasse reflects a tendency to disaffiliate from its perceived primary users rather than a disagreement over biological facts.

Hsiang-Chen Chi: *Being Understood in Empathy: A Philosophical Model*

The feeling of being understood is an important aspect of human experience that contributes to well-being, agency, and interpersonal connection. Although psychological research increasingly recognizes it as a significant dimension of empathy, philosophical discussions of empathy have paid comparatively little attention to the target person's experience of feeling understood. Existing theories often explain empathy in terms of how an empathizer comes to understand another's mental states, leaving unclear how the target person's feeling of being understood arises and what role it plays within empathic interaction.

To address this issue, I draw on Gillespie and Cornish's (2010) framework of perspective comparison. Their account suggests that understanding emerges through the comparison of

perspectives rather than through the accurate representation of another person's mental states. From this perspective, feeling understood arises when interlocutors engage in a reciprocal process of articulating, comparing, and revising their interpretations in relation to one another.

However, such a process presupposes certain epistemic conditions. Perspective comparison requires interlocutors to remain responsive to one another's interpretations while acknowledging the possibility that their own understanding may be incomplete or mistaken. I characterize this condition as reciprocal epistemic accountability.

Drawing on Bakhtinian accounts of dialogue, I argue that these epistemic conditions can be sustained only within a dialogic space. A dialogic space is a relational field in which meanings are jointly negotiated and remain open to revision. Rather than eliminating disagreement or uncertainty, such a space preserves them as conditions for inquiry and mutual understanding. Participants approach one another's perspectives without claiming complete access to them, and interpretive authority is continually negotiated rather than fixed.

Within this dialogic space, differences between interlocutors are neither erased nor overcome. Instead, they function as productive resources for empathic engagement. The possibility of misunderstanding remains present, yet interlocutors remain committed to examining and revising their interpretations in response to one another.

On this view, empathic understanding is not best conceived as the accurate grasp of another person's mental states. Rather, it is a co-constructed process through which interlocutors jointly clarify thoughts and feelings that may not yet be fully articulated. Consequently, the feeling of being understood is not merely an outcome of empathy but an integral component of empathic interaction itself.

Wednesday 1st July 17:00 — Gestures & Ostension (Kerkzaal)

Mirko Prokop: *Embodied Foundations of Ostensive Communication: A Gradualist Perspective*

According to an influential view, the flexibility and open-endedness of human communication relies on capacities for the production and interpretation of ostensive signals (Tomasello, 2008; Scott-Phillips, 2015; Levinson, 2022). The roots of this view derive from the work of Paul Grice (1957), who argued that, to communicative meaningfully (especially in novel situations), signallers must make their specifically communicative intent manifest to their audience, a process later referred to as ostension (Sperber and Wilson, 1995). In practice, signallers do this by means of behaviours such as direct eye contact, exaggerated movements, a tap on the shoulder, calling someone’s name or similar activities which, by openly addressing their audience, make it manifest that the signaller intends to communicate. Whilst ostension is widely thought to play a crucial role in the pragmatic interpretation of utterances, its developmental, evolutionary and cognitive basis remain a matter of debate (Csibra, 2010; Szufnarowska et al., 2014; Bar-On, 2013; Moore, 2016; Scott-Phillips and Heintz, 2023) According to the ‘natural pedagogy’ view, human infants are innately sensitive to a set of ostensive cues, including direct gaze, infant directed speech, and contingent responsivity (Csibra, 2010; cf. Csibra and Gergely, 2009).

On this account, infants recognise the communicative intention implicit in such behaviours and thus orient their attention to them, even before they gain access to the content of the intention. According to an alternative, ‘attention modulation’ view, infants’ heightened attention and sensitivity to other’s actions is not restricted to specifically ostensive cues, but may be elicited by

other behaviours which carry a social relevance, but which need not be communicative, e.g. someone's shivering (Szufnarowska et al., 2014).

Building on the latter view, I argue on both philosophical and empirical grounds that the understanding of ostensive signals (a) develops gradually on the basis of more general attentional biases towards others' bodily actions and (b) is not independent from a basic grasp of the intentional content of these signals.

On the philosophical front, I suggest that, especially in cases of embodied communicative interactions such as gestural or tactile engagements, it is very hard to conceptually separate the ostensive dimension of a behaviour from its intended content or meaning (cf. Moore, 2016, 2023). Since embodied communicative interactions are most pertinent in infancy, this suggests that the distinction between 'purely' ostensive signals (marking that someone intends to communicate) and their content (marking what they intend to communicate) is difficult to apply in an early developmental context. On the empirical front, I argue that this view is supported by two kinds of convergent findings. First, infants' selective attention to and intentionality understanding of others' actions which are related to their own action production and experience, suggesting that their heightened sensitivity towards others' actions is mediated by an embodied understanding of intentionality, which may include but is not restricted to ostensive behaviours (Gerson and Woodward, 2009; Liszkowski, 2018; Gredebäck, Gottwald and Daum, 2021). Second, comparative research on gestural communication in non-human great apes, which indicates that they can use and understand ostensive behaviours, but do so only for a relatively limited set of bodily gestures whose communicative function or 'meaning' is related to aspects of their own behavioural repertoire, which requires extended periods of interaction to develop (Pika and Fröhlich, 2019; Graham, Rossano and Moore, 2024).

Taken together, these findings support the idea that the understanding of certain behaviours as ostensive develops gradually from a more general attentional bias towards others' actions and intentions which is integrated with one's own action experience. I conclude that, from this perspective, the difference between human and non-human forms of communication – as far as the making explicit of communicative intent (i.e. ostension) is concerned – may be more a difference in degree than in kind.

Rory Harder: *Demonstratives Contribute Conventional Implicatures*

See PDF.

Eline Kuipers and Ludmila Reimer: *Rooting Iconic Co-Speech Gestures in Motor Representations*

During face to face communication, humans use a multitude of information channels to convey an intended meaning: speech, tone of voice and stress, facial expressions, body posture, and gestures. These channels are undeniably intertwined, and yet, it is still a puzzle how exactly they are linked on a cognitive level. To develop a cognitive model, one can turn to formal semantic approaches using grammar theories (Ebert et al., 2020; Schlenker, 2020) or to experimental methods investigating underlying cognitive functions (Kandana Arachchige et al., 2021). However, these approaches mainly focus on listeners and their comprehension of speech and gestures. So far, they do not offer a compelling theory of why and how gestures are (seemingly) effortlessly produced alongside speech by a speaker. Especially conic co-speech gestures (hereafter "gestures") are of interest, which are defined as co-occurring alongside speech and resembling the actions or objects that are talked about (McNeill, 1992, 2014), for instance tapping

your thumbs while saying “I’ll text you later”. In our framework, we postulate that the production of gestures depends on the mechanisms for intentional bodily action. More precisely, it captures the strong relation between motor representations and executable action concepts in the form of a psychological similarity space, extending Gärdenfors’ (2000, 2014) Conceptual Spaces Framework.

We argue that the seemingly unconscious and automatic production of gestures is based on the motor representations of the actions that the gestures express, which get activated through the executable action concepts that are simultaneously being communicated in speech. Concepts are thought to be mental representations corresponding to categories, where a category can be explained as a set of possible or conceivable exemplars, which are instances of these categories. Words strongly relate to the meaning of concepts, as humans often acquire concepts through word learning and express their thoughts, which are based on concepts, through language (Strößner, 2023). In turn, the semantic processing of action concepts at the lexical level is grounded in the processing of action concepts at the motor level (Ferretti & Zipoli Caiani, 2021).

Action concepts can be divided into executable and non-executable action concepts: Non-executable action concepts are purely observational, executable action concepts are concepts of actions one can, in principle, actually perform. Hence, while non-executable action concepts are based on one’s perceptual experiences of seeing others perform actions, executable action concepts are additionally based on one’s motor representations (Mylopoulos & Pacherie, 2017; Pacherie, 2011). Motor representations can be understood as “the representation of an action in the brain that is apt to determine the pattern of movements that the subject is going to perform in order to execute that action” (Brozzo, 2017, p. 233; based on Jeannerod, 1994, 2006). These patterns of movements are highly detailed, accounting for the temporal and biomechanical constraints of bodily movement (Shepherd, 2019). Executable action concepts are thought to enable one to select and activate matching motor representations in order to perform a specific action (Ferretti & Zipoli Caiani, 2019, 2021; Mylopoulos & Pacherie, 2017).

According to our model, motor representations are paths through a psychological similarity space, which we call the “motor space” (omitted). A similarity space is a multi-dimensional geometrical structure in which the geometrical distance between two points in the space represents the degree of (dis)similarity between two objects on given dimensions (Raffman, 2015). The dimensions of a psychological (or phenomenal) similarity space are dependent on an agent’s cognitive architecture (Gärdenfors, 2000). The motor space captures bodily parameters in bodily movements and similarities between specifications of all bodily parameters in an egocentric way (omitted). Within the motor space, the (dis)similarity between certain movements is captured by the distance between the representations of these movements.

We understand executable action concepts as corresponding to regions within the motor space (Gärdenfors, 2000, 2014). A region within a similarity space encompasses all exemplars that would be ascribed to the category that corresponds to this concept. Hence, within the motor space, there is a strong connection between executable action concepts and the motor representations of the actions these concepts capture. When one communicates, one accesses the concepts that are conveyed through one’s words and gestures, which are ultimately based on the (prototypes of) mental representations of one’s experiences. As motor representations determine the movements one is going to perform in intentional bodily action, these representations also lie at the basis of performing gestures. That is, when one wants to communicate the meaning of an executable action concept, a motor representation within the region that corresponds to this concept might get activated for gesturing.

To see if our model holds up in the real world, we need an approach that can verify it experimentally. Since we propose that actions and the gestures iconically representing them

relate to the same path through the motor space, they need to be measurably similar. However, measuring similarity is often difficult; for example, in lexical semantics, the semantic similarity of words is only indirectly measured by their co-occurence in written corpora and are often expressed as a single scalar within a two-dimensional space (Pennigton et al., 2014), thus conflating a multi-dimensional psychological similarity space of lexical semantics into a more tangible depiction. For our framework, we propose to also employ an indirect measure, namely to 1) motion-capture the movement of the hands during gesture production, 2) extract the corresponding shapes of the mean line the hands move through (gesture shape), 3) then compare it to the (also recorded and extracted) shapes of the actual activities represented by these gestures. We predict that the shape of a gesture is a truncated, simplified version of the shape of the corresponding action.

In conclusion, we claim that there is a strong, empirically discoverable relationship between iconic co-speech gestures, executable action concepts, and motor representations, which can be nicely modeled in a psychological similarity space, combined with Gärdenfors' (2002, 2014) Conceptual Spaces Framework. Furthermore, this model can explain the co-occurrence of a verbal and a bodily expression of an executable action concept in (at least) the form of iconic gestures accompanying speech.

Markus Werning, Ludmila Reimer, Thomas Wieder, Carla Zenk and Maria Spychalska: *How do iconic co-speech gestures contribute to the truth-conditions of assertions: A surprisal-based ERP investigation revealing N400 and late positivity effects*

See PDF.

Wednesday 1st July 17:00 — Generics & Narratives (Grote zolder)

Griffin Pion, Sophie Arnold, Elliot Schwartz, Julia Johnson, Eric Mandelbaum & Marjorie Rhodes: *Remembering Generalizations: Memory Mechanisms Underlying the Generic Recall Bias*

Humans have the capacity to form generalizations that extend beyond their direct experience. Those generalizations can be expressed linguistically in different ways, most prominently through generics (e.g., "Dogs bark"), which make claims about kinds without specifying exact prevalence information, and quantified statements (e.g., "All/Most dogs bark"), which explicitly encode proportional information.

A large body of work suggests that generics are a privileged way of expressing generalizations, and indeed may reflect a cognitively default way of forming generalizations (Leslie, 2007, 2008). First, developmental studies have shown that across multiple, unrelated languages, young children understand generics at a younger age than quantified expressions (Hollander et al., 2002; Mannheim et al., 2010; Gelman & Tardif, 1998; Gelman et al., 2016), and parents commonly use generics but not quantifiers in child-directed speech (Gelman et al., 2000, 2014). Beyond child behavior, Leslie et al. (2011) demonstrate that adults often treat quantified expressions as generics in reasoning tasks. One striking phenomenon supporting the view that generics express a cognitively default way of generalizing is the Generic Recall Bias (GRB): quantified statements are more likely to be misremembered as generic than generic statements are to be misremembered as quantified, both in adults and children (Leslie & Gelman, 2012; Gelman et al., 2016; Sutherland et al., 2015).

Despite the robustness of the GRB, its underlying causes remain unclear: the bias could arise during encoding, from the loss of quantifier information over time, or at retrieval when a stored representation is recalled. The present research addresses this question using a two-session

recall paradigm designed to dissociate these possibilities. Adults (N = 1,189) and young children (ages 4–9; data collection ongoing) completed versions of the same task; here, we focus on adults. Adults were randomly assigned to one of five language conditions (generic, all, most, some, or specific) and one of two memory conditions (immediate+delayed or delayed-only). All participants completed two sessions one week apart (fig. 1). In the first session, all participants completed a learning phase where they heard 16 statements about a novel social category (Zarpies) in their assigned language condition and 16 filler statements presented in the other language conditions. Participants then completed a distractor task. Those in the immediate+delayed condition were then asked to freely recall each statement from the learning phase; those in the delayed-only condition did not do any recall in the first session. In the second session, all participants were asked to recall the 16 Zarpies statements.

We identified three potential mechanisms that may lead quantified statements to be misremembered as generics. To accurately remember any given statement, you must (i) encode a representation into long-term memory, (ii) maintain the correct representation over time, and (iii) successfully retrieve this representation when prompted (Tulving & Thomson, 1973). The GRB could arise from any of these stages, leading to three hypothetical mechanisms:

(1) Encoding: When people hear a quantified statement, they encode a generic statement in long-term memory. (2) Retention: While people successfully encode the quantified statement into long-term memory, they forget the quantifier over time and only the generic statement remains. (3) Retrieval: When prompted to recall a quantified statement successfully preserved in long-term memory, people fail to retrieve the quantifier and report the generic. Our experimental design teases apart these hypotheses. If the GRB is due to lapses during encoding, then participants who misrecall a quantifier as a generic during an immediate memory test should also misrecall it one week later. In other words, encoding predicts a trial-level correlation between items misrecalled as generic. Conversely, if the GRB is due to lapses in retrieval, then misrecalling a generic immediately should not predict whether it is misrecalled after a delay. Finally, if the GRB is due to retention, we should expect participants to misremember quantifiers as generics more often when tested one week later than when tested immediately.

Results

First, we found evidence for the GRB in a wider range of quantifiers than previously documented: in addition to a GRB for “all” and “most” previously documented in adults, we also found a GRB for “some”. Second, and more significantly, we isolated the underlying mechanism of the GRB. We found no significant evidence in support of retention: the magnitude of the GRB was not moderated by memory condition (ps > .13; see figure 2). Instead, the results supported encoding most clearly: there was high trial-level correlation between recall at the first and second sessions (92% agreement; 4,470 matched trials out of 4,860). As further support for encoding, participants who recalled correctly in the first session always recalled correctly in the second session (3,579/3,579 trials). Finally, as additional evidence against retrieval, those participants in a quantifier condition who misrecalled a generic in the first session never recalled the correct quantifier in the second session (0/615 trials).

Overall, these findings provide converging evidence that the Generic Recall Bias arises primarily at encoding rather than during retention or retrieval. This suggests that generics communicate a privileged representation for storing kind-level information in memory, rather than a convenient linguistic shorthand at recall (see also, Gelman et al., 2016). In ongoing work, we extend our task to children to examine how this encoding bias develops across childhood. More broadly, the present longitudinal design offers a general framework for isolating the mechanisms underlying other asymmetric memory phenomena, such as the tendency to forget negations as affirmations rather than vice versa (Cornish & Wason, 1970; Maciuszek & Polczyk, 2017; Meyer, 1975). The

present study thus serves as a proof of concept that progress can be made on the mechanisms underpinning longstanding cognitive phenomena.

Elliot Schwartz: *How are Generics Defaults?*

An influential research program in cognitive science holds that generics are defaults. In other words, the meanings of generic sentences such as “Tigers are striped” are computed via the mind’s most basic mechanism of generalization (Leslie, 2007, 2008). The defaultness of generics is cited as an explanation for their ease of processing compared to quantified statements such as “All whales are mammals.” Compared to quantified statements, generics are understood earlier in development (Hollander et al., 2002), processed more quickly (Meyer et al., 2011), interpreted with fewer errors (Leslie et al., 2011), and remembered more accurately (Leslie & Gelman, 2012).

Here, I examine an underexplored question: how is this defaultness realized? What is it about generic statements that leads to the behavioral signatures just enumerated? I reject two proposals present in the literature, the explicitness account and the dual-process account. I then propose an alternative format account: generics are defaults because they are structurally simpler than quantified statements in respect to conceptual format.

Per the explicitness account, generics are easier than quantifiers because they involve cognitive processing performed without explicit instructions. As Leslie (2012) puts it: “If one wishes to interact efficiently with a system, and the system has a basic, default way of proceeding…then one need only issue an explicit instruction to the system if one wishes it to deviate from this default way of proceeding” (p.7). However, the fact that a system requires explicit instructions to perform certain operations does not entail that those operations are more difficult. For example, morphological processes frequently involve rules (e.g., add “-ed” to form past tense) that apply unless there is an explicitly memorized exception (e.g., “think” to “thought”). Nevertheless, morphological exceptions are generally faster to process than regular forms (since all exceptions are checked before the rule applies; Yang, 2016). The upshot: explicitness does not explain processing facts absent an account of implementation.

In other work, Leslie does suggest an implementation: generics are understood via fast, automatic, and effortless System 1 processes whereas explicit quantifiers are understood via slow, effortful, and rule-governed System 2 processes (Kahneman, 2003; Kahneman & Frederick, 2002; Leslie, 2007). On this dual-process account, when ascribing a property to category members, the generic generalizing mechanism applies automatically and must be actively inhibited to generalize using a quantifier like “all.” However, the contrast between these two systems is far from clear-cut: logical rule-governed operations like modus ponens, the supposed domain of System 2, are routinely carried out unconsciously and effortlessly à la System 1 (Quilty-Dunn & Mandelbaum, 2018; Reverberi et al., 2012). Moreover, it's unclear whether the dual-process approach is a genuine explanation of the facts. We know that some cognitive processes are less resource intensive than others, labelling these processes “System 1,” does not explain why this is so.

As an alternative, I argue that generics are defaults relative to quantified statements because generics employ a structurally simpler format at the level of conceptual representation. I assume that understanding a sentence involves constructing language specific representations (e.g., LFs) which are in turn used to form conceptual representations (e.g., LoT sentences; Knowlton et al., 2021; Quilty-Dunn et al., 2023). Conceptual representations may differ not only with respect to their content (e.g., DOG vs CARBURETOR) but also their format. For example, we might represent the meaning of “P and Q” using either a conceptual conjunction operator (i.e., P&Q) or a NAND operator (i.e., (P↓P)↓(Q↓Q)). In general, structurally simpler formats are easier to process

(Amalric et al., 2017). If a representation requires fewer steps to assemble, it should be faster to generate. Fewer steps also means less possibility for error relative to more complex representations. Finally, simpler representations require less cognitive resources and should therefore manifest earlier in development.

I propose that generics are understood by predicating properties of kinds (e.g., “Tigers are striped” is understood conceptually by predicating STRIPED of TIGERS) whereas quantifiers are understood by computing relations between sets of individuals (e.g., “All tigers are striped” is understood by computing a relation between the set of tigers and the set of striped things). All else being equal, predication is a simpler form of conceptual representation and so generics show the default processing signatures described above. However, the demands of set theoretic computation are lessened with restricted sets of concrete objects. Accordingly, quantified sentences should be easier to process in these conditions, a prediction borne out by experimental evidence (Gelman et al., 2015; Hollander et al., 2002).

Hamish Linehan: *Who chooses your past? Audiences, narratives, and environment*

Introduction

We understand ourselves through narratives. A successful autobiographical narrative should, in some sense, reflect our memories. Some have argued (Heersmink, 2020) that because our memories are distributed across objects, places, and people “who we are as narrative selves depends on and is partly constituted by a distributed network of environmental structures.” However, distributed narrative accounts have not fully captured the dynamic relationship between agents, memories, and environmental structures. This paper proposes that environmental structures do not merely store distributed memories. Instead, they actively shape those memories. Certain structures facilitate particular kinds of narratives. Narration requires the selective editing, omission, and emphasis of aspects of the personal past. Therefore, environmental structures can make agents narrate with a certain perspective, allowing the agent to experience their own memories with this perspective. In this way, environmental structures can cause, or be used by, agents to remember the past in particular ways.

What is a narrative?

Successful narratives have two key features. First, they are meaningful, allowing an audience to understand the narrator’s perspective at the time of the events. Second, they have emotional import, communicating the narrator’s present emotional evaluation of those events (Goldie, 2003). Both features are audience-relative. Narrators must anticipate what their audience will find intelligible or emotionally salient, selectively emphasising, omitting, or explaining aspects of the story (Goldie, 2012). During narration, audience feedback further shapes the account (Pasupathi et al., 2021). Through repeated rehearsal, narrators refine their ability to convey meaning and emotion, increasing the accessibility and familiarity of particular narrative framings (McAdams & McClean, 2013).

Affective Audiences

In many cases, narrators can choose their audience. Here, audiences can function as “affective scaffolds” (Colombetti & Krueger, 2015), helping to stabilise and amplify the emotional import of a narrative. Consider recounting an unusually wild weekend: told to a priest, the story may be structured around shame and regret; told to a friend, it may become humorous. Although the

remembered events remain constant, their meaning and emotional tone shift with audience expectations.

This highlights a dynamic relationship between memory, narrative construction, and audience response. The process resembles collaborative remembering, in which others help shape recall (Sutton, 2014), but differs in that narrators often actively select the audience for whom the narrative is constructed.

Consider an argument with a partner. Telling the story to the partner, especially if they believe you were at fault, requires engagement with their perspective. Recounting the same event to a partner-critical friend may encourage selective emphasis on details that minimise your responsibility. In each case, memories are re-experienced with different meanings and emotional valences, shaped by anticipated audience responses.

Repeated rehearsal of a narrative tailored to a particular perspective can result in “narrative railroading” (Osler, 2024). For instance, repeatedly telling an uncharitable version of an argument to a sympathetic friend may render that version the most accessible narrative for the narrator. When later attempting reconciliation, this sedimented narrative may undermine mutual understanding, having been constructed to deflect blame.

By anticipating audience expectations, narrators adopt particular perspectives that shape their narratives. Through rehearsal, these perspectives become entrenched, influencing how memories are re-experienced. Some aspects of the past become emotionally salient, others diminished or omitted altogether. In this sense, our narratives—and the audiences with whom we rehearse them—play a constitutive role in remembering. Memories should not be conceived merely as “the building blocks of narrative” (Heersmink, 2023); rather, narration is best understood as a dynamic feedback loop involving storytelling, audience expectations, rehearsal, and memory re-experiencing. Prescribed perspectives

Not all agents can choose their audience. In some contexts, individuals are compelled to make their narratives intelligible relative to prescribed perspectives imposed by their environment. Two illustrative cases are adults receiving an autism diagnosis and young men radicalised online.

Many autistic adults report long-standing difficulties meeting social expectations, often interpreting these struggles as personal or moral failures accompanied by shame. Receiving an autism diagnosis can provide a new narrative framework through which past experiences are reinterpreted. This shift is evident in first-person reports:

“The first 51 years of my life were absolute misery not knowing what I had, or why… I would think that I was a terribly wicked person because I couldn’t do many of the achievements that are ‘expected’ of ‘good’ people.” (Jones et al., 2001)

“Some of the personality traits which others led me to believe were faults or failings are not so and may be applied in ways which render them as assets.” (Lewis, 2016a)

“It was a bit like standing up in court and hearing the jury say: ‘not guilty.’” (Punshon et al., 2009)

Here, the emotional import of remembered experiences shifts: shameful memories are re-experienced with understanding rather than self-blame. Access to a neurodivergent perspective allows agents to reinterpret earlier difficulties, while also reshaping audience expectations and enabling more sympathetic responses from others. However, prescribed narrative scaffolding can also be harmful. This is evident in cases of online radicalisation among young men. Such individuals are often embedded in environments that exploit susceptibility to “narrative deference” (Byrne, 2025). Misogynistic online communities encourage members to reinterpret personal histories through narratives of grievance and victimhood, fostering what the UK Government

(2024) describes as a “preoccupation with reflecting on past instances of victimisation.” These narratives undermine alternative interpretive authorities and entrench a singular, adversarial understanding of experience, effectively railroading individuals into understanding their past through misogynistic frameworks.

Conclusion

These cases show that memories do more than merely support narratives, and that audiences are more than passive recipients of information. In some contexts, agents can choose their audiences, using them as affective scaffolds that shape emotional understanding and guide future action. In others, agents are subject to prescribed perspectives that impose particular emotional interpretations on experience. Together, these examples illustrate the dynamic interplay between memory, narrative construction, narration, and audience, underscoring the active role environmental structures play in shaping how the past is remembered.

Shawn Hsieh: *Grief as a Socially Embedded Process: A Mindshaping and Narrative Approach to Self-Reorientation*

This thesis fundamentally challenges traditional conceptualizations of grief as a simple, internal emotional response to death, proposing instead that grief constitutes a complex, socially embedded narrative process that extends far beyond bereavement to encompass diverse forms of loss, identity disruption and loss of life possibilities. Through integrating philosophical analysis with contemporary psychological insights, this work reveals how cultural contexts, social norms, and individual agency dynamically interact to construct and shape grief experiences. Traditional grief theories—exemplified by Kübler-Ross’s (2005) stage models and diagnostic criteria in DSM-5-TR (American Psychiatric Association, 2022)—conceptualize grief as a linear, time-bounded response to death that follows predictable patterns and requires “recovery” within socially acceptable timeframes.

This thesis systematically dismantles these assumptions by demonstrating that grief arises from fundamentally diverse loss experiences. Drawing on Ratcliffe’s (2022) phenomenology of grief and Cholbi’s (2021) concept of practical identity, I argue that what we truly grieve is not merely the fact of death or loss itself, but rather the disruption of life possibilities and relational structures that constituted our sense of self. When someone dies or a significant relationship ends, we lose not only that person but the entire practical identity constructed through roles, commitments, and imagined futures intertwined with that relationship. Grief thus emerges as an extended narrative process rather than a discrete emotional episode. As Goldie (2011) demonstrates, grief possesses an inherent narrative structure that unfolds temporally through patterns of meaning-making, identity reconstruction, and world-relearning.

This processual understanding explains why grief resists standardized timelines and varies dramatically across individuals—the depth of grief corresponds to how thoroughly the lost relationship or possibility was woven into one’s practical identity and life narrative. While acknowledging grief’s profound personal dimensions, this thesis reveals how grief experiences are fundamentally socially embedded through two complementary mechanisms: psychological constructionism and mindshaping processes. Psychological constructionism demonstrates that emotions are not fixed biological reactions but constructed psychological phenomena acquired through culturally embedded learning processes.

Following Barrett’s (2006, 2012) Conceptual Act Theory and Lindquist et al.’s (2015) language-emotion framework, I show how cultural contexts provide the linguistic resources, emotional vocabularies, and interpretive frameworks that literally construct what grief means and how it can be understood. Different cultures offer radically different grief concepts—some languages

possess rich grief terminology enabling nuanced emotional differentiation, while others collapse diverse loss experiences into broader categories. These linguistic differences are not merely expressive variations of universal grief; they fundamentally shape what grief experiences become possible within particular cultural contexts. Mindshaping processes extend this understanding by revealing how societies actively regulate and enforce normative expectations about appropriate grief behavior.

Drawing on McGeer's (2007, 2015) folk psychology framework and Zawidzki's (2008, 2013) social coordination theory, I demonstrate how cultural scripts, social expectations, and normative pressures don't merely provide interpretive resources but actively shape how individuals should experience and express grief. Through mechanisms including social amplification, emotion regulation, and culturally transmitted emotionology (Glazer, 2022), societies enforce specific grief timelines, intensity expectations, and expression norms that create powerful regulatory forces constraining authentic grief experiences. Together, psychological constructionism and mindshaping create what I term "Socially Embedded Grief"—a phenomenon where individual grief experiences emerge from continuous dynamic interaction between cultural frameworks, social regulation, and personal interpretation. Grief narratives unfold not in isolation but within social landscapes that simultaneously provide meaning-making resources and enforce normative constraints.

Recognizing grief's social embeddedness raises crucial questions about individual agency: if grief is socially constructed and shaped, how do some individuals maintain authentic experiences while others succumb to harmful social pressures? This thesis addresses this tension by examining how individual factors moderate social influences through the concept of narrative stability—the capacity to maintain coherent, self-authored grief narratives while navigating social expectations. Individual differences in attachment security (Bowlby, 1969; Ainsworth et al., 1978), personality traits (McCrae & Costa, 1987) particularly emotional stability and openness, self-identity clarity, and belief systems significantly influence narrative stability. Securely attached individuals with strong identity foundations and adaptive belief systems demonstrate greater capacity to selectively incorporate supportive social input while resisting pressures that contradict their authentic grief needs. These individual factors don't eliminate social influence but enable more balanced navigation between social expectations and personal authenticity. The thesis culminates in a critical distinction between two grief adaptation approaches that carry profound practical implications. The dominant social message to "move on" problematically assumes that healthy grief requires abandoning loss experiences, severing emotional connections, and returning quickly to previous functioning within predetermined timelines. This framework treats grief as pathological deviation requiring correction, creating internal conflict when authentic grief responses extend beyond socially acceptable boundaries.

Drawing on narrative therapy principles (White & Epston, 1990; White, 2007), I propose "moving forward" as a healthier alternative framework. Rather than demanding abandonment of what was lost, moving forward integrates grief into evolving life narratives while maintaining meaningful connections. Consistent with continuing bonds theory (Klass et al., 1996; White, 1988), this approach recognizes that adaptation involves not forgetting but rather reauthoring the significance of loss within ongoing identity development. Moving forward acknowledges that healing is not returning to previous states but developing new life structures that honor both past connections and future possibilities. This distinction carries practical significance: grief becomes not a problem requiring solution but a fundamental human experience deserving integration into ongoing life stories. By reframing grief as meaningful rather than pathological, moving forward creates space

for authentic emotional processing at individually appropriate paces, resisting social pressures for premature closure while maintaining social connection.

This thesis establishes grief as a complex, socially embedded narrative process that challenges reductive medical models. Understanding how cultural contexts and social norms construct and shape grief experiences reveals why universal diagnostic criteria fail to capture grief's genuine complexity. The distinction between “moving on” and “moving forward” offers practical implications: grief deserves integration into ongoing life stories rather than abandonment. We don't move on from grief; we move forward with it, weaving our experiences of loss into meaningful narratives of continued growth and self-reorientation.

Wednesday 1st July 17:00 — Inner Speech & Thought (Spiegelzaal)

Daniel Gregory: *What is Inner Speech? And what Does it Represent?*

Inner speech has become a well-recognized topic of interest in the philosophy of mind (see Gregory & Langland-Hassan (2024)). The question which has most interested philosophers is simply: What is inner speech? The same phenomenon as external speech, except that it is silent? An imagistic representation of speech sounds? Something else entirely? I propose that it is a unique mental state, not reducible to speech or to imagination.

Inner speech cannot just be silent speech. One reason is that inner speech does not involve the instantiation of concrete word tokens. No theory on the metaphysics of words allows that words can be tokened by sensory mental states. Such theories permit that instances of inner speech might involve imagistic representations of words, but not that they actually are words. (Thanks to [removed for review] for suggesting this to me.) The way we process inner speech also lacks some of the distinctive properties of the way we process external speech, such as speech segmentation, i.e., the division of a continuous stream of sound into apparently discrete units. Inner speech comes pre-segmented, as it were: there is no continuous stream that needs to be divided.

Inner speech also cannot be a kind of imagination—at least as imagination is usually understood. It is widely held that the contents of our imaginings—i.e., what they represent—are determined by our intentions (see Munro & Strohminger (2021: 11848, note 1), for proponents of this view). The view, typically attributed to Wittgenstein’s posthumous (1980), is articulated crisply by Fodor (1975):

"What makes my stick figure an image of a tiger is not that it looks much like one (my drawings of tigers don’t look much like tigers either) but rather that it’s my image, so I’m the one who gets to say what it’s an image of. My images (and my drawings) connect with my intentions in a certain way; I take them as tiger-pictures for purposes of whatever task I have in mind" (p. 191, his emphasis, as quoted by Munro & Strohminger (2021: 11850).

Munro & Strohminger (2021) complain that little argument has been given for the view, but an argument from elimination easily reconstructed from this quote provides some motivation for it. If the content of an analogue representation is not determined by a resemblance relationship, then the agent’s intention in producing the representation is simply the only remaining candidate.

The dominant theory of inner speech production posits a very different role for intentions. On this theory, the production of inner speech is connected to the process for monitoring our own physical actions. When we make physical movements, mental representations corresponding to the sensory feedback which should result from those movements form. For example, when motor commands to move your hand across your visual field issue, a representation forms of the visual experience which should result, i.e., the hand moving across the visual field. These mental

representations typically remain unconscious, but they facilitate rapid correction of our actions whenever there is a discrepancy between anticipated and actual feedback. It is thought that inner speech is produced when the physical process of speaking aloud is initiated but aborted almost instantaneously. A representation corresponding to the auditory feedback which would have resulted if the action were executed—the sound of oneself speaking aloud—nonetheless forms and, for reasons not yet understood, becomes conscious. (The literature is vast, but see, e.g., Tian & Poeppel (2012) and Grandchamp et al. (2019)).

If an instance of inner speech results from the initiation but then abandonment of an action, then its content cannot be determined by the agent's intention to represent something, precisely because the intention is never fully executed. So, inner speech cannot be a kind of imagination, at least if imagination is conceived of as an intentional sensory representation of some external object.

So, what is inner speech? It is clearly a sensory mental state, but what kind of content does it have? Channeling Anscombe (1957/2000), I argue that the content of inner speech results from what I will call 'abandoned action awareness' plus a process of deeming. We are directly aware of our intentions when we initiate but abandon actions, just as we are directly aware of our intentions when we actually perform actions. Even though an instance of inner speech results from the abandonment of an intention, we have direct access to that intention, which allows us to deem the inner speech to have the content which the initiated but abandoned speech act would have had, if it had been executed. The relevant intentions are intentions to express linguistic content. Thus, although it is a sensory mental state, inner speech has linguistic content (like external utterances), rather than auditory content (like auditory imagination).

One might object to the above and argue that inner speech is in fact a kind of imagination—a kind which actually complicates Fodor's analysis. That is, inner speech is a kind of imagination, the content of which is not determined by intentions. In fact, one might even challenge the standard interpretation of Fodor as claiming that intentions determine what images represent. After all, he says that images represent what we 'take' them to represent, and this 'taking' might sometimes involve deeming rather than intending (though this reading would involve downplaying his remark that '[m]y image … connect with my intentions in a certain way').

Perhaps. But the differences between inner speech and imagination, even where the latter involves deeming, run deep. The sensory experience in inner speech results from the abandonment, rather than the execution, of an intention. Inner speech is deemed to bear content which is derived from abandoned action awareness. And the content which inner speech is deemed to bear is content of a kind—linguistic content—which imagination is not otherwise thought to bear. For imagination, as usually thought of, represents objects via their perceptible properties.

On the face of it, then, inner speech is a genuinely unique phenomenon.

Daphne Bernués: *Why Inner Speech is Agentially Diminished*

See PDF.

Víctor Martín Verdejo and Marta Jorba: *Speaking in the Language of Thought*

Since the influential work by Willem Levelt and colleagues (Levelt 1989; Levelt et al. 1990), one central assumption in some accounts of inner speech is that it requires or presupposes a language of thought (LoT). On this conception, LoT is the linguistic medium needed to generate the input

to the entire speech production system. This view also aligns with a general conception of inner speech as involving the expression or vehicle of subpersonal amodal thought (e.g. Bermúdez 2003; Carruthers 2009, 2018; Prinz 2011). Moreover, recent developments see the LoT hypothesis as vindicated across a number of research fields and experimental paradigms in cognitive science (Mandelbaum et al. 2022; Quilty-Dunn et al. 2023). The natural presumption is that inner speech and associated capacities are clear candidates to be added to the list of LoT-supporting evidence. However, the question of whether inner speech really requires or presupposes a particular representational format deserves careful consideration. First, recent approaches have suggested that LoT is not a commitment of phenomenological or mechanistic approaches to inner speech or to the idea that we – sometimes or often – think in language (Vicente 2022; Kompa 2023). Secondly, fresh attempts to establish LoT on the basis of empirical evidence are in further need of clarification, and have been met with (sometimes staunch) resistance, with Deep Neural Networks (DNNs) standing as well-supported alternatives (LeCun, Bengio, & Hinton 2015; Millière 2024).

The goal of this paper is twofold: (i) first, to provide conceptual clearing of the ground on the connection between inner speech and the LoT; (ii) second, to outline a functional, empirically informed answer to the question of whether inner speech presupposes or in any way underscores a LoT. Regarding (i), articulation of the exact connection needs to take into account the characterizations of LoT and inner speech on the market. The LoT hypothesis, introduced by Jerry Fodor (1975, 2008), posits that thinking occurs in a mental language (or Mentalese) which is structurally similar to natural language.

We will introduce the view that LoT and competing hypotheses are better seen as affording integrated multi-level explanations of cognitive phenomena in a given domain. A multi-level approach strengthens the plausibility of the LoT by integrating explanations across (roughly Marrian) levels in ways that meet the demarcation constraint (LoT and non-LoT formats can coexist and their explanatory import identified), the specificity constraint (different LoTs may be postulated depending on the specific domain of application), and the empirical constraint (empirical evidence must be possibly gathered in favor of, or against a particular LoT when set against alternatives). Unlike the property-based approach favoured by Quilty-Dunn et al. (2023) this account requires the proper delineation of explanatory levels, and in particular the functional, the algorithmic, and physical realization levels. Inner speech (IS), on the other hand, broadly refers to the phenomenon of talking to oneself silently, one prominent kind of inner speech being verbal thinking or thinking in words. Yet the connection between IS and LoT needs to heed a diversity of characterizations of IS as well.

The commitment to LoT disintegrates, for instance, if IS is considered to be only, or essentially, an introspectively accessible kind of experience keyed to a particular natural language. If restricted to a personal and conscious level, IS would be compatible with any underlying mechanism that results in that experience. However, IS has been also defined by appealing to the speech production mechanism, where inner speech is the result of a multi-stage process of linguistic articulation, going from the formulator level to the semantic/syntactic and auditory-phonological levels. In Levelt's model, LoT structures must be retrieved at the conceptualization stage as part of a multilayered process which gradually incorporates features of outer speech. From this perspective, LoT has a distinctive functional role to play at the "preverbal" stage.

In this case, however, the proposal has to address the question of "overdetermination", given that it would imply the postulation of two different mechanisms for explaining how verbal thinking acquires its contents– a LoT and a speech production route. On a more liberal interpretation of Levelt's model, IS only requires episodes that substantially engage the speech production system, so that episodes of thinking could result without requiring a LoT (Kompa 2023). The landscape of available characterizations of IS is however rich (Gregory and Langland-Hassan 2024). The

present suggestion is that only if IS has associated a more specific functional role, can the connection with LoT structures be meaningfully stated. In relation to (ii), therefore, we will examine a functional account of IS along these lines, where a particular cognitive function is associated with IS experience or IS production mechanisms. From this perspective, a commitment to LoT is possible depending on the representational format that is required in cognitive processing to fulfill a target cognitive function.

The idea that IS is associated with cognitive functions is widely accepted at least since Vygotsky's work (1934). The assessment of whether these functions are LoT-supporting requires minute demarcation of the functions at hand. An initial hypothesis to be considered is that LoT is more plausible in some cognitive tasks (e.g. problem solving, memory recall) than others (e.g. motivation, self-regulation). Even then, the ultimate criterion for a significant commitment is empirical and degree-like. We may find cases in which inner speech functions are likely subpersonally realized in a LoT, as well as in alternative nonsymbolic representations, including associations or computational DNN systems, which have been successfully applied to system-1 processing, language translation, or image recognition. The resulting approach would steer a middle course between those who tend to presuppose LoT to be a condition on IS – as suggested by Levelt's influential account – and those who tend to presuppose that IS and the corresponding notion of "thinking in language" is a phenomenon to be reductively explained in terms of format-neutral natural language capacities.

Yizhi Li: *Beyond Memory: Goal-Structured Dynamics in Spontaneous Thought*

Recent research on spontaneous thought has been shaped by two converging trends: an emphasis on memory as the core underlying mechanism and a growing interest in its dynamics—the way one thought gives rise to the next within a temporally extended spontaneous thought stream.

This paper argues that these trends, when combined, reveal a fundamental gap in current theorizing. While memory-based accounts can explain the content of spontaneous thought, they provide an incomplete account of its dynamics because they overlook a distinct mode of organization: goal-structured dynamics, in which cognitive control actively shapes how the stream of thought unfolds. The content/dynamics distinction. Following the dynamic framework of Christoff et al. (2016), I treat spontaneous thought as a temporally extended phenomenon whose explanatory target includes not only what is thought (content) but how successive thoughts are selected and connected (dynamics). The same content can appear in very different streams: one governed by associative drift, another embedded in deliberate planning. Thought-transition dynamics constitutes a dimension of the explanandum distinct from, and irreducible to, content. As I will argue, it is precisely this dynamic dimension that reveals the insufficiency of memory accounts.

Memory accounts and their shared limitation. A recently emerging family of memory accounts shares a foundational commitment: specific memory mechanisms lie at the core of spontaneous thought. By reducing spontaneous thought to well-studied memory processes, theorists aim to leverage our understanding of memory to illuminate the less-understood phenomenon of spontaneous thought. Despite their heterogeneity, these accounts converge on a systematic blind spot: cognitive control is excluded from the story of how spontaneous thought unfolds. I briefly review several representative memory accounts: the SWR-trigger hypothesis (O'Callaghan et al., 2021), the consolidation-reflection view (Wamsley, 2019), the pattern-completion account (Mills et al., 2018), and the unconstrained memory framework (Mildner & Tamir, 2019, 2024). Those that address dynamics characterize thought transitions as fundamentally associative. Goals as

content versus goals as organizers. The unconstrained memory framework, the most sophisticated and comprehensive memory account, can accommodate two well-known types of spontaneous thought content: goal-related and future-oriented content. According to the framework, current concerns can modulate associative retrieval: pending goals have heightened accessibility within the associative memory structure (Klinger, 2013), and pre-formed plans can enter consciousness via control-free associative retrieval as "memories of the future" (Cole & Kvavilashvili, 2019). However, there is a crucial distinction between a goal's entering consciousness, which is a matter of content, and a goal's organizing the thought stream across multiple transitions, which is a matter of dynamics. An activated goal might simply drift away via association, or it might be elaborated into a goal-directed process, e.g., making a plan during spontaneous thought. These possibilities cannot be distinguished by looking at content alone.

Associative dynamics versus goal-structured dynamics.

I introduce a distinction between two modes of thought-transition dynamics. In associative dynamics, next thought selection is local: the immediate predecessor thought, together with modulatory weights, determines the next thought via the strongest available association. In goal-structured dynamics, next thought selection is non-local: a maintained goal representation, which I term an organizing goal, exerts a stable influence across multiple transitions, with each successor thought selected by its instrumental relevance to advancing the goal. Instrumental relevance is the selection criterion; means-end coherence is the observable structural pattern this criterion produces.

I argue that this distinction is irreducible. The objection that goal-structured dynamics can be reconstructed as a strong modulatory bias within the associative framework fails because the two modes implement structurally different selection principles. Modulatory bias changes which candidate wins a local associative competition, but cannot introduce a non-local representation that governs selection across multiple steps. The distinction receives independent support from the model-based/model-free distinction in reinforcement learning (Daw et al., 2005).

Goal-structured dynamics in spontaneous thought. I then demonstrate that goal-structured dynamics is a recurring feature of spontaneous thought through two empirical routes.

First, I address a pervasive ambiguity in the literature: most studies operationalize "planning during mind wandering" as goal-related, future-oriented content without assessing whether the thought stream exhibits the instrumental relevance structure characteristic of actual planning. I then present evidence for a spontaneous planning hypothesis, suggesting that genuine planning, which involves cognitive control, occurs during spontaneous thought: future-oriented mind wandering (a type of spontaneous thought) involves more inner speech and structured sequences (Stawarczyk et al., 2013), is sensitive to executive resources and working memory capacity (Smallwood et al., 2009; Baird et al., 2011), co-activates the default and frontoparietal control networks (Spreng et al., 2010), and refines personal goals toward greater concreteness (Medea et al., 2018).

Second, conversational simulation (or imagined interactions (Honeycutt, 2003)), imagining anticipated or past communicative encounters, provides a second route. I argue that simulating a conversation requires control to maintain a dialogic frame as an organizing goal that constrains transitions: successor thoughts are selected by relevance to the simulated exchange, not by associative strength. Supporting evidence includes the dominance of social content in spontaneous thought, with participants frequently reporting that they put themselves in others' shoes (Mar et al., 2012; Diaz et al., 2013), the prevalence of imagined interactions in diary studies (Honeycutt et al., 2014), and the neural overlap between dialogic inner speech and default

network regions, which are reliably activated during spontaneous thought (Alderson-Day et al., 2016).

A dual-driver framework. These arguments motivate a hybrid account of spontaneous thought in which memory and cognitive control are both important. Memory supplies representational materials and associative pathways; control—automatically recruited in at least some cases—maintains organizing goals that shape the stream. The two interact dynamically: an associatively cued concern can recruit control when elaborated into, for example, planning, and a control-guided sequence can dissolve into associative drift when the goal loses salience. I propose that the alternation between an associative exploration regime and a goal-structured pursuit regime is a functional feature of spontaneous thought. Finally, I argue that the field's recent methodological innovations (e.g., the free-association semantic tasks, thinking-aloud protocols) are insensitive to goal-structured dynamics due to their conceptual limitations, and call for a reconceptualization and refinement of these methods.

Wednesday 1st July 17:00 — Imagination & Memory (Voorkamer)

Sofia Pedrini: *Remembering in Someone Else's Shoes: Vicarious Memory and Empathy*

Vicarious memories (VMs) are "recollections people have of salient life episodes that were told to them by another person" (Pillemer et al., 2015, p. 234). Unlike episodic memories (EMs) of one's personal past, VMs are formed through conversation, written narratives, or social media sharing (Thomsen et al., 2025). Recent psychological research has shown that VMs share core phenomenological qualities with EMs—vivid mental imagery, emotional intensity, and similar physical reactions—and play similar adaptive roles in identity formation, self-understanding, and decision-making (Pillemer et al., 2015; Pond and Peterson, 2020; Panattoni and Thomsen, 2018).

Despite growing psychological interest, VMs have received little philosophical attention. The phenomenon raises interesting questions: How does the sense of reliving differ when we remember an event we witnessed versus receiving testimony about it? How does this affect the sense of self in VM? Do VMs involve some sort of autonoesis?

In this paper, we argue that we can better understand the intensity of VMs, the emotional charge they carry, and their adaptive roles if we consider the encoding stage of VMs not simply involving testimony or communication, but (also) an act of empathy.

Traditional accounts of empathy—whether theory-theory, simulation theory, or direct perception theory—focus primarily on empathizing with another person's current mental states during direct interaction, such as understanding their movements, behaviors, gestures, expressions, emotions, and actions (Gallagher, 2020). Memory, however, is a higher-order mental state compared to these directly observable phenomena. When we encode someone else's memory, we are not witnessing the original event itself (which occurred in their past, before our encoding), but rather we empathize with another person through her testimony or narrative about that event.

We build on Edith Stein's phenomenological account of empathy to shed light on the form of empathy that is involved in VM's encoding stage. For Stein, empathy proceeds in three steps: (1) emerging awareness of another's mental state, (2) "fulfilling explication", where we are drawn into this state, and (3) comprehensive objectification of the mental state (Stein 1989, p. 19). The second phase is imaginative and it plays a crucial role: we join the other's perspective by imaginatively adopting their standpoint (Dullstein 2013). We "co-live" the other's experience, being guided by their perspective while remaining aware that this mental state is not primordially our

own—it possesses “co-originality” (Konoriginalität). The imaginative second phase allows us to transcend perceptual awareness and enter the other’s experiential perspective.

We propose that VM encoding involves this higher-order form of empathy. When someone shares a significant life episode with us, we do not merely receive informational content; we engage in an empathic process that involves: (1) direct perception of the other person’s emotional expression, (2) to imaginatively enter their perspective on the remembered event, sharing (con-living) the affective dimension of their experience, and (3) objectifying and integrating this empathically grasped content into our own memory system. The emotional charge is not merely “about” the story we’re told, but emerges from sharing the narrator’s affective state toward the remembered event.

This empathic dimension explains at least two puzzling features of VMs. First, it accounts for their phenomenological similarity to EMs: vivid imagery and emotional intensity arise from empathic engagement with the narrator’s own recollection. Second, it clarifies why the narrator-listener relationship matters: VM intensity correlates with the narrator’s importance for our identity because empathic engagement deepens with emotional connection.

The emotional investment required for empathic engagement also explains why not all testimony produces VMs. For a narrative to become encoded as VM rather than mere semantic knowledge, we must enter empathy’s imaginative phase, being drawn into and guided by the narrator’s perspective. This arguably requires both the narrator’s capacity to convey their experience and the listener’s empathic engagement.

This account has broader implications for understanding the intersubjective constitution of memory and identity. If VMs are formed through empathy, then significant portions of our autobiographical landscape—which shapes our sense of self—are constituted not just individually but through empathic engagement with others. The memories that orient our decisions and self-understanding include not only what we have experienced, but what we have empathically shared with significant others.

Christopher Jude McCarroll, Ying-Tung Lin and Paloma Muñoz Gómez*: Memories of Fiction, Mindshaping, and the Porous Self*

Autobiographical memories are, in some sense, what make us who we are (Rowlands 2017; Schechtman 1994). The experiences we have over the course of a life are retained in memory, and form the building blocks for our sense of self and our continuity through time. The idea that the experiences that constitute the self are maintained in memory is much too simple, however. For one, memory is a constructive capacity, which draws on many sources of information (Bartlett 1932; Roediger & DeSoto 2015; Schacter 2012). It is also a capacity that is tightly connected to, perhaps even continuous with, imagination (Michaelian 2016). Moreover, our memories are not individual and internal records of the personal past, but may be distributed across the physical and social world (Sutton 2002; 2010). In this way, our selves too might spread out into the world and include material and social elements incorporated into our autobiographical topography (Heersmink 2018; Fabry 2023). Material and social elements seep into our porous selves.

One striking example of this is vicarious memory. Whereas personal memory involves the recollection of events in the personal past that were experienced firsthand, vicarious memory involves recollections of events that happened to other people (Pillemer et al. 2024). These vicarious experiences are shared with us through narratives about the past, and these stories allow us to imaginatively enter into the lifeworlds and experiences of others. It is thought that such vicarious memories serve similar functions to personal memories (Pillemer et al. 2024). This is

one way in which the self is porous, incorporating aspects of experiences that were originally undergone by another person. One's autobiography is enriched with elements of vicarious experiences. In this way, there is a broadening of our autobiographical repertoire and our sense of self.

Some researchers want to push the boundaries of the self even further, to include not merely vicarious memories of experiences that were had by perhaps close others (friends, family members etc), but memories of events portrayed in the fiction of films, books, and other media (Marsh & Yang 2020). Adopting a functional approach to memory systems, Marsh and Yang (2020) argue that, just like memories of events in the personal past and vicarious memories, memories of fiction should also be considered as occurrences of event memories. On this view, memories of fiction are phenomenologically and functionally similar to memories of personally experienced events and hence can be integrated into our autobiographical records (Marsh & Yang 2020). According to this line of thought, memories of fiction are similar to vicarious memories, contributing to the general functions: identity, directive, and social□of autobiographical remembering (Reese 2025).

It is this fictional turn in connection to memory and the self that we examine in this paper. Drawing on existing research, which adopts a functional approach to autobiographical memory, we first outline the ways in which memories of fiction might be important elements of our autobiographical repertoires. This existing work leaves us with important questions about the role of engaging with fiction and the ways in which memories of fiction might relate to the self. Engaging with fiction is often thought to be a way of improving social cognition (Kidd & Castano 2018; Hutto 2007): it enables us to become better and more accurate mindreaders, allowing us to predict and explain the behaviour of others through the accurate attribution of mental states.

In this paper we adopt a different perspective. We suggest that part of the function of fiction is that it serves a mindshaping role (Zawadzki 2013; cf. McGeer 2007; Mameli 2001). Fiction is one way in which we are provided with virtual models, which describe and sanction examples of culturally appropriate normative and moral behaviours. Our memories of fictional events are in this way a form of vicarious learning, shaping us to enter into the normatively appropriate mental states and behaviours relative to our specific cultural groups. We outline this claim in detail and demonstrate how engagement with fiction encourages enculturation and impacts the self by providing us with virtual role models that we can approximate. In this mindshaping view, our memories of fiction lead us to become more cognitively homogenous, sharing mental states and emotions with others. Through virtual role models and templates of appropriate behaviour, engagement with fiction and our memories of fiction allow us to become more like one another. In some ways this mindshaping role of fiction can imprison us in certain roles or identities, enforcing stereotypes and encouraging us to adopt behaviour and mental states considered appropriate to members of particular groups (Wolf 2025; cf. Peters 2019). This is especially true in the case of master plots or narratives (Lindemann Nelson 2001; Fabry 2025), which can, for example, provide stock examples of how to live a life and outline what kind of trajectory it should have (McLean & Syed 2015). However, as we demonstrate, engaging with fiction can also provide us with new role models with which to imaginatively engage, encouraging us to shape our own selves and break out of certain ways in which our identities might be constrained (Kind 2024).

Virtual models in fiction can provide us with guidance on how-to: how to organise our emotions, how to read those of others, how to judge the appropriateness of certain behaviours. They also provide us with explanatory value. We make sense of our actions and those of others in terms of the roles we and others play, or are expected to play. Virtual role models provide us with

information and have a sense-making function in our lives. In this way, memories of fiction are crucial aspects of our autobiographical repertoires and play crucial roles in shaping our identities.

Andreas Arslan and Jonathan Kominsky: *Incoherent mental imagery: Where imagination and episodic memory diverge*

The idea that episodic memory and imagination are related is not a recent one: Hume, in A Treatise of Human Nature, claimed that memories and imagined events cannot be reliably distinguished based on either their constituent 'simple ideas' or their overall structure (Hume, 2003). It is only the 'superior force and vivacity' of memories that allows us to tell them apart from pure fancy. In the last two decades, research on episodic memory has begun to call even that graded distinction into question: The notion that episodic memory and imagination are functionally related, in particular that episodic memory is a specific instance of a more general mental capacity to simulate events, has gained increasing traction in cognitive science and neuroscience (Schacter et al., 2007; Addis et al. 2007; Addis, 2018), as well as philosophy (De Brigard, 2014; Michaelian, 2016; Mahr, 2020), and seems to be substantiated by clinical findings (Hassabis et al., 2007). So, was Hume mistaken, when he asserted that 'ideas of the imagination' are 'fainter and more obscure' – does it turn out that representations in episodic memory and imagination are on par in terms of vividness? Here, drawing on our own new empirical work on imagination, we will argue that there are certain consequential dissimilarities between the construction of imagined scenes and the encoding of experiences in memory.

We will conclude with a sketch of some theoretical arguments that challenge the hypothesis that episodic memory and imagination-based representations are of a similar format. One of the essential characteristics of episodic memories is that they tend to change or degrade over time. How forgetting (Anderson & Hulbert, 2021; Radvansky et al., 2022) and the progressive abstraction or 'semanticization' of episodic memories (Nagy et al., 2020) unfold is a subject of ongoing research and debate. But when comparing imagination and episodic memory, it is important to keep in mind that fresh memories of even trivial events initially are often highly vivid and comprehensive (Talamini & Gorree, 2012) – which might be due to a tendency to 'promiscuously' encode details (Hardt et al., 2013). By contrast, a set of five preregistered experiments we recently conducted, indicates that imagined scenes are usually highly 'incomplete' from the outset, at least as far as the representation of spatial relations is concerned. In our experiments, participants had to complete a 'contradiction detection task' (Arslan & Kominsky, 2025), in which they were asked to imagine a short text as vividly as they could. All vignettes we presented to participants contained an objectively contradictory description of spatial relations. For instance, an open door is described to be 10 steps across from a window at the beginning of the text; later, there is suddenly a couch in that very same place. Through a series of questions, we tested whether participants had noticed the contradiction.

Our first experiment replicated the central finding of Arslan and Kominsky (2025) in a considerably larger sample, showing that only a minority of participants succeeds at detecting the contradiction. The four follow-up experiments investigated the phenomenon of incomplete or incoherent spatial representations in imagined scenes more comprehensively, by systematically modifying the texts participants had to imagine. In Experiment 1, we observed that only 35 out of 100 participants detected the contradiction in a short (192 words) vignette that was very explicit in its descriptions of spatial relations. Experiment 2 replicated this finding (detection rate of 38.5%), and in Experiment 2.1 we observed that 'isomorphic' vignettes that contain the same error as the original text, but are entirely different in terms of setting ('cave' and 'clearing,' as opposed to the original 'living room'), are associated with similarly low detection rates. Experiment 3 showed that rephrasing spatial descriptions to make them somewhat more indirect significantly reduces

detection rate (11%), $\chi 2(N=151,2)=11.90$, $p>0.01$. Experiment 4 varied the instructions, specifying more precisely how to imagine/read the text ('imagine vividly', 'read carefully', 'POV', 'bird's eye view'), yet detection rates were the same for all conditions, $\chi 2(N=180,3)=0.73$, $p=0.87$. Experiment 5 tested whether adding descriptions of salient causal relations would increase the detection rate. It did not, $\chi 2(N=150,2)=3.33$, $p=0.19$. Importantly, in Experiments 2.1 – 5 participants were asked how vividly they had 'seen' the scene they imagined in their mind, on a scale of 1 ('Not vividly at all') to 5 ('Very vividly'). These subjective ratings were consistently high (between M=3.93 and M=4.21) but consistently failed to correlate with success at detecting the contradiction.

These results imply that imagined scenes are at no point as similar to on-line perception as relatively recent episodic memories. While our empirical work is confined to representations of spatial relations, we want to point out a few other characteristics of imagined scenes that might set them apart from memory representations. First, there are reasons to believe that representations of time should be just as rudimentary as that of space: Whereas the encoding of potentially irrelevant occurrences (that later are pruned away) during the formation of an episodic memory might serve as an indicator of duration, it is unclear how anything comparable can be accomplished by way of pure imagination. In other words: What would it mean to 'simulate' waiting for five hours? Second, imagination, in many instances, is a top-down process that starts with an abstract intention (e.g., trying to imagine a house) that initiates the generation of concrete mental imagery. The opposite –an ambiguous 'imagined percept,' triggering recognition – is difficult to conceive of: 'What is this supposed to be? Oh, it appears I'm imagining a house.' The formation of memories, however, clearly involves such perception-driven, bottom-up processes. Third, an episodic memory (ideally) eventually converges on one coherent representation of a specific event ('I still remember what grandma's house looked like on that day in 1981.'). Conversely, it is often useful to imagine multiple contradictory drafts of a fictitious scene (e.g., many different houses), which are never put together into a coherent final version and quickly forgotten, once they have served their immediate purpose.

Ariel Gonçalves, Kourken Michaelian and Antônio Jaeger: *Fundamental dissimilarities between the event-related potentials of remembering and imagining and their implications for continuism*

Over the past two decades, considerable evidence from neuroimaging, behavioral, lesion, developmental, and aging studies has revealed important psychological and neural similarities between episodic memory and episodic imagination (for a review, see Schacter & Addis, 2020). An influential theoretical position known as "continuism" has emerged from this literature (Perrin, 2016), positing an identity between the capacity to remember one's past and the capacity to imagine one's future. As presented by Michaelian (2016), continuism involves the "same system" claim, according to which a unified neural system underlies both capacities. The same system claim was supported by the neuroimaging evidence accumulated by the episodic simulation research program. While the distinct theories that are encompassed by this research program differed in their specifics, they shared the core idea that newfound neuroimaging results pointed to a previously unrecognized neural similarity between memory and imagination. The more radical proponents believed, more specifically, that they had discovered that a single, unified brain network was responsible for both episodic memory and imagination. As we will show, this may have overstated the extent of the similarity, leading to an unjustified confidence in the conclusiveness of the same system claim. A blind spot that may have contributed to this unjustified confidence derives from the fact that the evidence produced by the research program has a critical limitation: the neuroimaging studies relied predominantly on fMRI and PET.

These techniques have coarse temporal resolution, with task epochs in representative studies commonly ranging from 4 to 16 seconds in fMRI and early PET work averaging over 30-second blocks. This temporal granularity risks blurring potential distinctions between remembering and imagining, as it can overlook dynamics that unfold faster than the tool's detection threshold. Thus, drawing conclusions exclusively on the basis of data thus gathered risks mistaking an artifact of temporal smoothing for an overlap between remembering and imagining. In the work reported here, our objective was therefore to investigate whether the continuist same system claim appears to be well supported when examined using a technique that allows for millisecond-level temporal precision. This tool is the event-related potential (ERP) technique, which has been extensively used in memory research (for a review, see Kwon et al., 2023).

The present study included nineteen participants (13 female; M age = 22 years), who completed a modified recognition paradigm. The experiment comprised 16 blocks, each with two phases. In the first phase, participants viewed 7 AI-generated images for 2000 ms each. In the second phase, 14 word-pairs were presented individually for 2000 ms. Half of these word-pairs corresponded to images shown in the first phase (e.g., "boy–shouts" for an image of a boy shouting), while half were unrelated to any encoded images. Participants were instructed to form a vivid mental image based on each word-pair and press a key once such an image was achieved. Crucially, no reference to the previously viewed images was made during these instructions. After each trial, participants indicated whether their mental image was similar or identical to an image seen in the first phase. The remember condition included trials where word-pairs matched encoded images and participants reported both vividness and similarity; the imagine condition included trials where word-pairs had no corresponding image and participants reported vividness but no similarity to a previously shown image. The parietal old/new effect, a well-established ERP marker of episodic recollection in recognition memory studies, was present for the remember condition but absent for imagination. This is demonstrated by a significant main effect that emerged in the 500–800 ms ($F(1, 18) = 7.90$, $p = .011$) and 800–1100 ms ($F(1, 18) = 10.9$, $p = .003$) windows, with the remember condition eliciting greater positivity than the imagine condition. Topographic analyses confirmed that these effects were particularly pronounced in the relevant electrodes.

This dissociation suggests that the remembering and imagining conditions may engage different neural processes during this critical time window associated with episodic retrieval. The uniqueness in the ERP signature of memory indicates that confidence in the same system claim was unjustified once temporal dynamics are examined with greater precision. As noted earlier, the experimental results of the episodic simulation research program led some researchers to conclude that a single brain network was responsible for both memory and imagination. This became a common reading of that literature and was adopted by both psychologists and philosophers. But there were, in fact, researchers within the episodic simulation program who, from the outset, put forth alternative readings that did not endorse the more radical same system interpretation. We side with these readings and propose a deflationary view. On the deflationary view, the same system claim is treated merely as a live hypothesis, one that is plausible but not one on which there is consensus or one that is self-evident. This deflation does not, however, amount to an outright rejection of the same system claim. It need not, then, mean the end of continuism.

The deflationary view does, however, suggest that continuism would benefit from revision, since treating the same system claim as a live hypothesis requires criteria for assessing whether empirical findings corroborate or undermine it. Clarifying what it means for the neural systems underpinning two mental capacities to be one and the same would provide such criteria. One way of achieving this clarification is to understand continuism as a mechanistic thesis (Camillo, 2025) according to which episodic memory and imagination manifest the same cognitive function if they are realized by similar interactions among similar neural entities. Continuism, then, does not

crumble with the deflation of the same system claim; it continues to serve as a valuable framework for understanding experimental findings, now equipped with criteria that allow it to be assessed against the shifting empirical landscape.

Wednesday 1st July 17:00 — Intention & Free Will (Bovenkamer)

David Barack and Daniel Burnston: *Plans, Planning, and Intentions*

Planning is central to agency. Bratman has proposed a well-known account of the relationship between planning and agency where plans and intention stand in distinctive relations as part of a larger agential architecture [1]. Here, we critique Bratman's approach from the perspective of reinforcement learning (RL), focusing on two of his main claims: the formation of intentions (i) precede planning and terminate deliberation and (ii) prompts a switch from reasoning about ends to reasoning about means.

We contend that RL, a widespread influential computational approach in cognitive science, puts pressure on both theses. In RL, agents make choices and observe their outcomes. Any mismatches between outcomes and expectations are used to update the expected outcomes for subsequent decisions. Agents learn the value of actions via these updates to use in selecting actions in the future. In model-based RL, agents also learn the probability of transitioning between states, such as the probability that state S2 occurs if the agent performs action A in state S1. Planning involves using this model to simulate a series of actions off-line. Given the transitions and associated probabilities, the simulations allow the agent to calculate the expected outcome of potential actions. In this framework, action values are always considered in the context of a model of the environment. Different environments involve distinct contingencies that change the expected value of these actions. To make a decision, the environmental model is used during those simulations, calculating the expected outcomes as part of the decision process. Planning behaviors are explained by the RL framework. For instance, in multi-stage tasks [2], agents make a series of choices, where choices in the first stage are probabilistically related to what options will be available in the second stage. If agents are using a model to make decisions via simulation, their first-stage choices will reflect the overall reward history for that choice, the contingencies of the second-stage choice, and the probability of the transition between the stages. That is, the initial choice will be considered in the context of the probability of ending up in the second-stage and which options will then be available. Changing the probabilities in the action-state transition or modifying the value of the second-stage choice outcomes will be reflected in the probability that the agent makes a particular choice at the first stage. The influence of first-state outcomes, transition probabilities, and second-stage rewards is a kind of result frequently observed in complex choice tasks [3-6]. We argue that sensitivity to environmental transitions and to differences between actions depending on states, outcomes, and transitions is evidence that agents are planning their actions through simulation in a model.

Given the explanatory power of this framework, we suggest the following thesis: (P-RL): Planning is simulation of outcomes from future decisions using a model of the environment. This (P-RL) thesis captures the idea that decision making is part of planning in virtue of the processes involved in simulation. Bratman's commitments are challenged by (P-RL). Bratman's first claim is that intentions terminate deliberation. On (P-RL), planning is deliberation via simulation, involving beliefs about probabilities of transitions between states of the environment due to actions and their desirable outcomes. But then, intentions at best are simultaneous to planning and deliberation. The second claim is that intentions prompt a switch from thinking about ends to thinking about means. On (P-RL), the expected values of options are computed by simulating outcomes using the agent's model, which comprises series of state-action transitions that probabilistically lead to outcomes. Since these are representations of means on which decision-making operates, the (P-RL) thesis implies selection of means and ends simultaneously. This

includes simulation at different grains, such as choosing to get a drink and choosing which arm angles and trajectories to select to pick it up, where such fine-grained actions are gathered together as 'options' that enter into the simulation [7]. P-RL is a major shift from standard philosophical thinking about plans. Besides undermining Bratman's account, this shift has significant implications for cognitive ontology, the norms of planning and action, and cognitive control. We end by outlining these briefly.

Lilian O'Brien and Svetlana Vetchinnikova: *The Side-Effect Effect and "Intentionally"*

Joshua Knobe's (e.g. Knobe 2003) work has led many philosophers to accept the "side-effect effect" (SEE): SEE: Foreseen, negative, but unintended outcomes are judged by the folk to be intentional actions – the folk concept of intentional action is systematically affected by normative judgments.

A key source of disagreement about SEE concerns what exactly respondents aim to convey when they agree that the CEO of Knobe's vignettes "intentionally harmed the environment". To shed light on this, we examined "intentionally" in the Corpus of Contemporary American English (COCA, Davies 2008), where it occurs 7,511 times (ca. 7 per million). Our analysis suggests that "intentionally" predominantly occurs in negative contexts involving acts of violence, deception, theft, oppression, and emotional harm. Its occurrence in positive contexts is comparatively rare. This presents, we argue, a challenge to SEE: if "intentionally" is used to attribute the performance of intentional action, it should appear with greater frequency in positive contexts. This raises the question of why respondents agree that the CEO intentionally harmed the environment. Two promising hypotheses about the use of "intentionally" are considered, one involving negative semantic prosody and the other involving meaning change. These hypotheses promise to explain why respondents find it natural to judge that the CEO intentionally harmed, but such judgments are not, contrary to SEE, judgments that the CEO performed an intentional action.

Berke Aydas and Onurcan Yilmaz: *Intuitive and Reflective Foundations of Free Will and Scientific Determinism*

Belief in free will has been examined across philosophy, psychology, and cognitive science, yet there remains little consensus regarding its cognitive foundations. Within free will research, psychology have primarily sought to address three core research questions: (1) in the absence of free will, what changes in human behavior (Vohs & Schooler, 2008); (2) in what sense does it exist (Bargh, 2008); and (3) why people believe in free will (Clark et al., 2014). However, this body of work has often not directly addressed the underlying cognitive processes that give rise to belief in free will and endorsement of scientific determinism (defined here as the view that human actions are fully caused by prior states of the world and governed by lawful, non-random processes). To our knowledge, no prior research has systematically examined the cognitive foundations of belief in free will and scientific determinism. A substantial body of evidence suggests that human judgment and decision-making rely on two broad classes of cognitive processes: intuitive, fast, and affectively driven processes, often labeled System 1, and reflective, slow, and deliberative processes, often labeled System 2 (Kahneman, 2011).

These dual-process frameworks have been widely applied to domains such as reasoning (Evans, 2011), moral judgment (Bago & De Neys, 2019), and belief formation (Fugelsang & Thompson, 2003). Despite the prominence of dual-process theories, little research has directly examined whether beliefs about free will and scientific determinism are differentially grounded in intuitive versus reflective cognition. Addressing this gap is theoretically important, as such beliefs may not merely reflect abstract philosophical commitments but may instead emerge from the same

cognitive mechanisms that shape everyday judgment and reasoning. The present research program applies Dual Process Theory to investigate the cognitive origins of belief in free will and scientific determinism. Across three experiments, we experimentally manipulated intuitive and reflective processing and measured their effects on endorsement of free will and determinism using the Free Will and Determinism Plus Scale (FAD–Plus; Paulhus & Carey, 2011). To further examine the robustness and generalizability of these effects, we adopted a cross-cultural approach, comparing Turkish and Canadian samples (in line with open science practices, we registered all of our experiments in this research: https://osf.io/p7gnz/overview?view_only=4f5d2f2d795642f0957e1685e4a5342e).

Hypotheses We tested three primary hypotheses. First, intuitive processing induced through time pressure and emotion priming was expected to increase belief in free will and decrease belief in scientific determinism relative to control conditions. Second, reflective processing induced through debiasing training was expected to decrease belief in free will and increase belief in scientific determinism. Third, under time pressure, participants were expected to report higher belief in free will and lower belief in determinism compared to control trials. Cultural differences were examined exploratorily.

Method and Results

Experiment 1 Experiment 1 tested these hypotheses in a large sample of Turkish adults (N = 747) using a mixed design. Participants were randomly assigned to reflection via long debiasing training, intuition via emotion priming, and time pressure (with an embedded within-subject time-delay manipulation after time pressure), or control conditions. Results indicated that reflective training reduced endorsement of both free will and scientific determinism, while time pressure increased belief in free will and decreased belief in determinism. Effect sizes were small for determinism ($\eta^2 = .02$) and medium for free will ($\eta^2 = .09$). These findings led to the emergence of a novel proposal—the Reflective Doubt Hypothesis—suggesting that reflection may introduce doubt toward initially held intuitive beliefs, reducing endorsement of both metaphysical positions rather than selectively increasing determinism. Experiment 2 Experiment 2 (N = 344 Turkish university students) employed a similar design but removed the within-subject time-pressure manipulation and added both active and passive control conditions. Reflective debiasing training and emotion priming both reduced endorsement of scientific determinism compared to active control ($\eta^2 = .03$), though no significant effects were observed for belief in free will. Experiment 3: Experiment 3 (N = 727) tested the Reflective Doubt Hypothesis by examining whether an analytical thinking manipulation reduced endorsement of free will and scientific determinism relative to active and passive control conditions in a Canadian student sample. No significant differences were observed between conditions for either belief, with all tests yielding nonsignificant results (ps > .05). As an alternative analytic approach, we examined intellectual humility, a construct conceptually related to reflective doubt and epistemic openness. However, no condition differences emerged on this measure, suggesting that the impact of reflective processing on belief change may depend on specific contextual or procedural features.

Discussion

Across three experiments and two cultural contexts, the present findings suggest that reflective reasoning does not straightforwardly increase endorsement of scientific determinism, as initially hypothesized. Instead, when effects emerged, reflection was associated with a general weakening of endorsement of both free will and determinism. This pattern supports the Reflective Doubt Hypothesis, according to which reflective processing may introduce epistemic uncertainty toward initially intuitive metaphysical beliefs rather than selectively promoting one worldview over another. Importantly, these results help clarify the psychological relationship between free will and scientific determinism. Rather than functioning as simple opposites, beliefs about free will and

determinism appear to be jointly sensitive to changes in reflective engagement, with reflection sometimes reducing confidence in both.

This finding aligns with prior evidence that lay beliefs about free will and determinism are often compatibilist in nature and suggests that reflection may operate on belief certainty rather than belief content alone. Intuition-based manipulations, particularly emotion priming, produced less consistent effects across studies, despite passing manipulation checks. This variability highlights potential boundary conditions for when intuitive processes shape belief endorsement. More broadly, the present work contributes to the growing literature on the cognitive framework of belief systems by demonstrating that core beliefs are not fixed commitments but can be sensitive to cognitive processing modes. By integrating dual-process theory with research on free will and determinism, this research provides a process-oriented account of how such beliefs are formed and revised.

Judith Carlisle and Ryan Mokhtari: *Possible Futures: Free Will and the Stories We Tell Ourselves*

Clinical research suggests that treatment engagement and persistence improve when people see themselves as agents whose choices, efforts, and self-regulation matter (Amdie et al. 2022; Baumeister et al. 2009). By contrast, experimental attempts to manipulate belief in free will through brief determinism primes have produced weak and inconsistent behavioral effects (Vohs & Schooler 2008; Genschow et al. 2020, 2023). This paper argues that this apparent mismatch has two sources. First, small, short-term belief manipulations are unlikely to produce robust changes in long-term behavior. Second, and more importantly, belief-manipulation studies target abstract commitments about free will, whereas clinical practice relies on local, practical forms of agency that are graded, context-sensitive, and responsive to sustained intervention (Haggard 2002; Nataraj et al. 2020). We propose a framework for articulating agency in clinical contexts that supports motivation and responsibility without presupposing controversial metaphysical claims.

Wednesday 1st July 17:00 — Representation & Explanation (Kleine zolder)

Basile Jeannet: *No Evidential Test for Naturalistic Theories of Content*

Naturalistic Theories of Content (NTCs) aim to account for representational content in non-intentional, non-semantic terms. Varitel semantics (Shea, 2018) and informational teleosemantics (Neander, 2017) are two recent proposals offering contrasting views about content. On what grounds should we adjudicate among NTCs? One influential response in the literature is to put the NTCs to an evidential test, i.e., to a comparison between the content ascriptions predicted by the NTC and those found in our best scientific explanations.

I assess two common ways to put NTCs to an evidential test: by comparing them with (1) explicit scientific content ascriptions, and (2) reconstructed content ascriptions. I highlight issues related to both approaches. I argue that (1) does not provide unambiguous content ascriptions and that (2) relies on the very theories it is meant to test. In the absence of an evidential test, philosophers should think about new ways to adjudicate among NTCs. The notion of an evidential test captures a common strategy when discussing NTCs.

Philosophers interpret convergences between NTCs' claims and scientific content-ascriptions as confirming the NTC. "If the scientists' ascriptions concur with the ones our theory entails, this can help to confirm the theory" (Neander, 2017, p. 115). Conversely, divergences are taken to signal an issue for the theory. This approach approximates what I call an evidential test for NTCs. An evidential test is a comparison between the content ascription predicted by the NTC in a given

situation and that assumed by scientific theories in the same situation. I define what it means for ST (scientific theory) to constitute an evidential test for an NTC:

i. The NTC predicts that a mental state S has content X in circumstances C

ii. ST claims that S has content Y in circumstances C, and would ascribe the same content if C were to occur again (robustness requirement)

iii. ST is a reliable source of evidence about content (reliability requirement)

iv. ST is independent from NTC (independence requirement)

v. NTC is confirmed if, given C, X = Y, vi. NTC is disconfirmed if, given C, X ≠ Y.

"Independence" means that ST's content ascriptions should not depend on NTC. For example, a change in NTC should not modify any of ST's content ascriptions.

Could explicit scientific content ascriptions form an evidential test for NTCs? In explaining behavior, scientists mention explicit representational contents, which they attribute to internal vehicles or to steps in computational explanations. Why not test NTCs against them? I raise two problems. First, whether scientific explanations involve representations or not is not transparent. Psychologists and neuroscientists use highly context-sensitive expressions, like "response", "representation", "being elicited by", "being about", "recognizing", "signal", "being sensitive to". Those terms may or may not denote genuine representational talk. Moreover, neuroscientists, psychologists and philosophers tend to be uncertain about how to apply the concept of representation, suggesting that the very concept of representation is imprecise (Favela & Machery, 2023). The second problem is that explicit content ascriptions vary across and within studies. For example, studies on A2 cells in the auditory system of the moth tend to provide inconsistent content ascriptions. While one study mentions "bat's echolocation calls" (Ratcliffe et al. 2009), another refers to a "neural representation of bat predation risk" (Goerlitz et al., 2020) and yet another mentions ultrasound (ter Hofstede et al., 2013). These variations show that explicit content ascriptions are inconsistent across studies, and therefore fail to meet the robustness requirement (ii).

If explicit scientific content ascriptions do not constitute a good evidential test for NTCs, then reconstructed content ascriptions might. Moving away from explicit ascriptions to reconstructed ascriptions is common in the literature (Neander 2017, Shea 2018). These reconstructions typically insist on the explanatory role of content. I suggest that some reconstructions are subject to a circularity risk: they are partly determined by the theory they are meant to test. These reconstructions then fail to meet the independence requirement (iv). I focus on Neander (2017). Neander reconstructs the explanatory role of content by analyzing neuroethology explanations of how the toad detects "wormlike" stimuli (Camhi, 1984). Since these explanations describe toads' visual T5-2 neurons as causally sensitive to worm-like stimuli rather than to prey, Neander concludes that the content playing an actual explanatory role is worm-like stimulus, not prey. Yet this conclusion cannot be reached from these considerations alone. It needs the additional premise that content implies a causal sensitivity to the represented condition (Neander, 2017, p. 156). This additional premise does not derive from the scientific explanation, but from Neander's theory of content. The reconstruction therefore fails to meet the independence requirement.

What should we do then? One option is to give up on the idea that content properties are objective. Deflationism (Egan 2020, 2025) suggests that content properties are ultimately determined by research interests. From the deflationist's perspective, the divergences among NTCs stem naturally from the plurality of research interests. Deflationism does fit some of the observations made above about inconsistent content ascriptions and imprecision in representational talk within the scientific literature. However, there remain strong reasons to stick to objective

representations, not least because mainstream psychology routinely appeals to accuracy conditions in its explanations (Burge 2010). Another option would be to give up on the idea of evidential tests. Theories can be assessed through a great variety of means, including internal coherence, external coherence, or simplicity. The idea of testing theories of content against scientific explanations remains appealing, if only in the minimal sense that NTCs should not contradict our best scientific explanations. Finally, it must be noted that while Neander (2017) is subject to the circularity risk, Shea (2018) is more cautious and makes no explicit appeal to evidential tests. Further work should expand on the different ways that NTCs relate to scientific explanations.

Anastasia Garbayo: *What comparative biology reveals about the explanatory role of representations*

The status of representations remains a central and contested issue in philosophy of biology and cognitive science. In recent decades, enactive, dynamical, and organism-centered approaches have convincingly demonstrated that many forms of adaptive, flexible, and apparently goal-directed behavior can be explained without invoking internal representations. By emphasizing real-time organism-environment coupling, distributed physiological dynamics, and the self-organizing properties of living systems, these approaches have significantly reshaped our understanding of biological cognition and challenged representationalism as the default explanatory framework. However, an open question remains whether this non-representational strategy can be generalized across the full range of biological systems, or whether there are principled limits to such explanations grounded in biological organization itself.

This paper argues that representational explanations provide genuine explanatory understanding only at specific points here non-representational accounts reach their limits, and that thee limits coincide with a biologically significant organizational transition associated with the emergence of neurons. The argument is developed withing the framework of the representation-hungry problem introduced by Clark and Toribio (1994). According to this view, representations should not be treated as default explanatory posits but are warranted only when behavior cannot be adequately explained in terms of continuous dynamical coupling between organism and environment. The central philosophical task, then, is not to decide whether representations exist in general, but to identify when and why representational explanations become explanatorily indispensable rather than merely convenient or metaphorical. To address this question, the paper adopts a comparative, empirically informed approach drawing on two biological case studies that differ in organizational complexity. The first concerns complex nutritional decision-making in the slime mould Physarum polycephalum, a non-neural organism. Experimental work shows that Physarum can flexibly regulate its intake from multiple food sources with different nutritional compositions, dynamically adjusting consumption in accordance with its internal nutritional state (Dussutour at al., 2010). This behavior involves the integration of multiple environmental variables, context-dependent choice, and apparent optimization over time, and is frequently described using cognitive vocabulary such as “decision-making”, “preferences”, and “nutritional targets”.

Despite its sophistication, I argue that nutritional regulation in Physarum can be fully explained in non-representational terms. The relevant explanatory resources include distributed physiological processes, growth dynamics of the plasmodial network, and differential absorption rates, and continuous feedback loops between organism and environment Crucially, the behavior does not depend on learned associations, localized memory structures, or the selective modification of internal states based on past outcomes. No internal states need to be interpreted as encoding relations between nutrients or representing nutritional goals. In this case, representational posits do not increase explanatory power or understanding, even though the behavior might prima facie

appear to motivate them. The second case concerns associative learning in the box jellyfish Tripedalia cystophora, an organism with a decentralized nervous system. Recent experimental work demonstrates that box jellyfish can learn to associative specific visual patterns with mechanical collisions and subsequently modify their swimming behavior to avoid obstacles (Bielecki et al., 2023). This learning is mediated by localized neural plasticity within the rhopalia and exhibits key features such as stimulus specificity, sensitivity to temporal relations, and persistence across changes in environmental conditions. I argue that these features cannot be adequately explained by appeal to global biochemical changes or undifferentiated organism-environment coupling alone. Instead, they require positing internal states whose functional role is to selectively stabilize relations across time and guide behavior in contexts that are not currently present. In this case, representational explanations are not optional redescriptions but are required to account for how learning is organized, maintained, and generalized beyond immediate interaction. The contrast between these two cases reveals a qualitative evolutionary transition in biological organization. Neurons introduce new organizational capacities, including structurally localized plasticity, outcome-dependent modification of internal states, and the ability to stabilize behaviorally relevant relations across time.

Together, these capacities undermine purely non-representational explanations and generate genuine representation-hunger. The paper's main contribution is to clarify how and when representations provide understanding in biological explanation. Representations do not merely label behavioral regularities; they explain why certain patterns of behavior are stable, selective, and counterfactually robust across time and circumstances. By grounding representational necessity in comparative biology and organismal organization, this account offers principled criteria for determining when non-representational explanations suffice and when representational posits are explanatorily required.

Filippo Murabito: *Beyond the Mark of the Cognitive: An Explanation-Centred Heuristic for Plant Cognition*

While humans are typically regarded as the paradigmatic cognitive agents, growing empirical evidence suggests that cognitive processes may extend far beyond the animal kingdom. Recent advances in plant biology indicate that plants exhibit complex and flexible behaviours that, when observed in mammals, are often taken to warrant cognitive attribution (e.g., Segundo-Ortin et al., 2026; Segundo-Ortin & Calvo, 2023; Trewavas, 2015). Reported examples include decision-making at both root (Hodge, 2009) and shoot levels (Gruntman et al., 2017), anticipatory responses (Novoplansky, 2016), communication (Runyon et al., 2006), kin and species recognition (Bilas et al., 2024), memory (Vyse et al., 2022), and various forms of learning (Gagliano et al., 2014, 2016). This expanding body of work, frequently discussed under the label "Plant Neurobiology" (Baluška & Mancuso, 2007; Calvo, 2016), has prompted many philosophers, biologists, and cognitive scientists to reconsider whether cognition might extend to organisms lacking nervous systems (Lyon et al., 2021; Lyon & Cheng, 2023). The resulting debate is both conceptual and methodological, concerning how cognition should be characterised, which theoretical principles ought to guide its investigation, and how experimental practices might support meaningful comparisons across taxa. More broadly, these questions bear directly on attempts to reconstruct and understand the evolution of cognition across the tree of life.

A common rationale for interpreting plant behaviour as cognitive appeals to complex forms of information processing (e.g., Novoplansky et al., 2024; Shoot et al., 2025). The guiding intuition is that plants detect, integrate, store, and exploit information to regulate adaptive responses to changing environments, and that such capacities plausibly support cognitive ascriptions. However, this strategy faces two fundamental challenges. First, the concept of "information"

remains underspecified in these contexts: different philosophical accounts of information carry distinct theoretical commitments and explanatory consequences (Fresco, 2022; Piccinini & Scarantino, 2011). Second, the nature of cognition itself remains deeply contested: there is no consensus that information processing is either necessary or sufficient for cognition (e.g., Adams, 2018; Di Paolo et al., 2017; Varela et al., 2016), and no resolution in sight. As a result, debates over plant cognition often stall at the level of competing conceptual frameworks, rather than generating cumulative epistemic progress (Colaço, 2023; Lee, 2023).

This paper addresses these difficulties in two stages. First, it clarifies and critically assesses prominent philosophical conceptions of information under which plants might be said to process information, focusing on Dretske's (1986) natural information, Floridi's (2011) environmental information, and Bateson's (1979) difference-making account. While plausible cases can be made that plants process information in each of these senses, the analysis shows that none of these conceptions, taken in isolation, suffices to ground claims about cognitive status. Whether such information-processing capacities warrant cognitive attribution ultimately depends on how cognition itself is characterised, and on resolving the broader impasse surrounding the mark of the cognitive (Rowlands, 2010).

In response, the paper advances a methodological shift from metaphysical boundary-drawing to explanatory practice in cognitive science by introducing the Explanatory Parity Heuristic (EPH). According to the EPH, when a paradigmatic cognitive phenomenon is best explained by positing a certain class of informational relations or organisational structures, and when an explanation of plant behaviour likewise invokes structurally similar relations playing similar explanatory roles, these explanatory similarities provide defeasible but non-trivial support for extending analogous explanatory inferences to the plant system with respect to the targeted capacity. The EPH thus reframes the question of plant cognition as an explanation-centred inquiry, focusing on the roles that informational posits actually play in successful scientific explanations rather on metaphysical criteria.

The EPH does not attempt to redefine cognition or to establish information processing as either necessary or sufficient for cognitive status, thereby sidestepping the longstanding impasse over the mark of the cognitive. Instead, it articulates a general methodological principle: similar explanatory roles warrant similar explanatory inferences, unless independently motivated asymmetries can be identified. On this way, the central question becomes not whether plants satisfy pre-theoretical intuitions about cognition, but whether the best explanations of their behaviours rely on organisational structures that perform the same kind of explanatory work as those invoked in paradigmatic cognitive systems. Furthermore, the support delivered by the EPH is explicitly defeasible and capacity-relative. It targets specific explananda rather than global claims about the cognitive status of plants. Resulting inferences remain provisional, since apparent explanatory parity may be undermined by alternative accounts that explain the same behavioural patterns without invoking information-processing organisation in an indispensable way and without corresponding explanatory loss. Accordingly, the EPH serves as a pragmatic methodological guide, bolstering confidence in extending cognitive-style explanatory inferences precisely when informational organisation proves explanatorily indispensable in the same manner as in established cognitive domains.

Thus, the paper makes two primary contributions. First, it clarifies the explanatory implications of leading conceptions of information as applied to plant systems. Second, it develops and motivates the Explanatory Parity Heuristic as a principled tool for distinguishing cases in which informational talk is merely rhetorical and dispensable from cases in which it contributes substantively to explanation. Applied to representative examples from contemporary plant science, such as root risk sensitivity under temporally variable nutrient conditions (Dener et al., 2016), the EPH provides a framework for strengthening (or weakening) confidence in attributing specific cognitive

capacities to plants, while avoiding reliance on contested a priori definitions of cognition. More broadly, the proposed approach provides a general template for evaluating cognitive attributions in non-neural organisms by grounding them in comparative explanatory practice, helping to move the debate on plant cognition from conceptual deadlock toward empirically testable progress.

Thursday 2nd July 09:00 — Keynote (Kerkzaal)

Ira Noveck: *Reconfiguring Figurative Language*

Paul Grice's Theory of Meaning and his innovative program led him to analyze words, utterances composed of words, and exchanges composed of utterances and to carefully consider how words were *used* in utterances as well as in conversational contexts. Ultimately, he provided a framework designed to expose the inner workings of communication; central to it was his notion that communicating is about *connecting with another mind*. To come up with predictions, early experimentalists developed what became known as the *Standard Pragmatic Model* (the SPM), according to which the literal meaning of an utterance was processed until a maxim was violated, at which point the addressee (the participant) would effortfully come up with a new pragmatic reading. Investigations based on this model, which relied to a great extent on Reaction Times, led to doubt about Grice's account. In this talk, I review work – on metaphor, irony and idioms -- in part to show that this early (SPM-centered) experimental approach led us down a path that did not do Grice justice and in part to provide an update on the processing of these three figures, as inspired by an Experimental Pragmatic approach.

Thursday 2nd July 10:45 — Plenary Symposium (Kerkzaal)

Nick Allott, Jane Dilkes & Diana Mazzarella: *Understanding figurative language: functions, attitudes, and social meaning*

The nature of figurative language and its relationship to literal language remain the subject of ongoing debate. How do figurative uses of language differ from their literal counterparts, if they do? Why do speakers choose figurative expressions over available literal alternatives? Do figurative uses of language serve specific communicative functions? Which attitudinal meanings are associated with different types of figurative language use? More broadly, how does figurative language contribute to the expression of stance, evaluation, and social alignment? By examining a range of figurative phenomena from complementary theoretical and empirical perspectives, the symposium aims to advance our understanding of the cognitive, communicative, and social dimensions of figurative language.

Nicholas Allott: *The literal/figurative distinction and the functions of figurative uses of language* (Joint work with Mark Textor)

According to radical pragmatics the meaning of lexical words is routinely modulated in use. Since many such modulated uses are non-figurative this casts doubt on the literal/figurative distinction (Wilson & Carston 2007): deviation from lexically encoded meaning is the norm, not an exception. But the literal/figurative distinction matters to accounts of communication because literal uses carry information that figurative ones don't: if used literally "John is a gorilla" tells us about gorillas, but not when it is meant as a metaphor (Searle 1979). We propose a Non-Conformity View of Non-Literal Use: literal uses of a word are made with the intention to continue a 'tradition' of using a word; figurative uses are made with the intention to deviate from the tradition (Allott & Textor 2026). In this talk I outline our view and show how it can shed light on the functions of figurative uses of language.

Jane Dilkes: *The function of figurative language*

This talk focuses on metonymy as a fundamental figurative category. To investigate the functions of metonymy in different contexts, the study analyses discourse in six online groups organised into three categories relating to group purpose. Individual and group metonymies are compared across these groups. The analysis combines qualitative examination with computational methods. There are differences in prevalence and function of metonymy between the three categories of groups. While in some groups individual metonymies are not prevalent, in other groups they are used extensively for rhetorical purposes. In some groups the function of group metonymy is shown to be significantly different from non-figurative use of the same term; in other groups use of group metonymies over time aligns with predictions from social identity theory. Across contexts, metonymy appears to support movement in relation to conventional usage, or wider social norms. Group metonymies inherently represent the stance and focus of a group, including the distance of a group from wider norms.

Diana Mazzarella: *Irony as attitude expression: a developmental perspective*

Verbal irony is often distinguished from other forms of figurative language in that it is primarily *interpretive* rather than descriptive (Wilson, 2006). The primary goal of an ironical speaker is not to describe a state of affairs, but rather to express a dissociative or critical attitude towards the proposition literally conveyed. Consequently, a developmental account of irony comprehension should explain how children come to recognise the implicit and dissociative nature of ironical attitudes and distinguish ironical utterances from mistakes or lies. In this talk, I present findings from a series of experimental studies we conducted with 4- to 8-year-olds to shed light on the challenges young interpreters face and the socio-cognitive capacities that support their developing understanding of ironical uses of language.

Parallel Sessions

Thursday 2nd July 4:30 — Symposium: Reshaping Normalcy: Mindshaping and Injustice in Mental Health (Kerkzaal)

This symposium examines whether mindshaping views of mental interpretation (McGeer, 2007, 2021; Zawidzki, 2008, 2013) offer a more flexible and potentially emancipatory theoretical framework. On this view, the primary function of mental interpretation is not to track independently constituted mental states, but to regulate agents' behaviour and reasoning in line with the socionormative expectations embedded in folk-psychological concepts (e.g., beliefs, desires) and other interpretative tools. This emphasis on the socially scaffolded and norm-governed nature of mental interpretation shifts attention toward the social practices through which the limits between the pathological and the "normal" are established and contested (Ballesteros et al., 2025; Russell, 2024; Zawidzki, 2024).

The first contribution combines the mindshaping framework with an embodied account of uptake, explaining cases of epistemic injustice and wilful hermeneutic ignorance in doctor–patient communication as distinct forms of breakdown in the regulative spiral of mutual understanding. The second examines cases of hermeneutical injustice driven by impostor or distorting concepts that undermine self-understanding, arguing that mindshaping is better suited than realist, mindreading-like accounts to provide an ameliorative account of the legitimacy of agents' self-concepts. Finally, the third turns a self-critical eye on the mindshaping framework, highlighting a

pernicious idealization about first-person authority at its core that may hinder its ability to properly handle cases of self-demeaning or unjust self-attributions in mental health.

Kathleen Murphy-Hollies & Jodie L. Russell: *Mindshaping and Uptake: Illuminating Epistemic Injustice and Wilful Hermeneutic Ignorance*

When communicating with others, there can often be significant social and personal stakes to the conversation going well. We therefore want to express ourselves in the ways which are most likely to be received and understood by others, or, in other words, to receive uptake from others. The clinical encounter between doctors and patients is one such situation where the stakes are high (miscommunication might compromise patient care), and where both patient and doctor (normally) seek a successful social exchange. Nevertheless, exchanges in the clinical encounter can go wrong. In this talk, we demonstrate how the processes of mindshaping (Zawidzki, 2013) capture not only the ways that both parties aim to be understood, but also the harmful ways in which this can be unsuccessful. Mindshaping is the view that social agents understand one another through mutually conforming to, or subverting, folk-psychological norms, and characterises the process of understanding as something that unfolds dynamically over time (what Andrews, 2015, terms "the folk-psychological spiral"). While mindshaping gives a compelling account of how social understanding succeeds and, at other times, fails (due to mismatched norm conformity), it is a primarily descriptive account of social cognition. However, we think it can additionally help us identify and explain normative cases of harmful, unsuccessful social communication. Namely, cases of epistemic injustice (Fricker, 2007) and wilful hermeneutic ignorance (Pohlhaus, 2012). These are cases where more powerful knowers fail, or refuse, to acknowledge the epistemic tools of more marginalised knowers, and this enables powerful knowers to unjustly misunderstand, misinterpret or ignore particular knowledges. However, it is difficult to differentiate between cases of epistemic injustice and wilful hermeneutic ignorance. We suggest that coupling mindshaping's account of understanding with Whitney's (2018) embodied account of uptake, developed from Merleau-Ponty (2013), can illuminate this. According to Whitney, individuals are said to give uptake when they are moved by the embodied and affective expressions of another. In cases of epistemic injustice, interlocutors fall short of being appropriately moved by the other, and uptake is not afforded. According to a mindshaping framework, marginalised knowers are then not admitted into the regulative spiral of mutual understanding and norm setting. In cases of wilful hermeneutic ignorance, individuals actively and knowingly inhibit themselves from being moved by another, and thus the folk-psychological spiral becomes one-sided (only one person puts in the work to understand and be understood) or collapses entirely. On an embodied account of uptake, wilful hermeneutic ignorance becomes evident in situations where speakers and listeners refuse to embody norms of uptake which would signal one's attempt to communicate, and so such phenomena can be very much 'felt' by interlocutors. In turn, this explains patient's visceral testimonies of not being 'seen' and 'understood' by doctors. In clinical contexts, doctors may fail to recognise testimony that falls outside dominant medical norms and narratives for understanding, or they may wilfully refuse to exercise the skill inherent to the 'spiral' of mutual understanding in folk psychological practices.

Víctor Fernández Castro & Miguel Núñez de Prado-Gordillo: *Whack-A-Mole: Mindshaping, Impostor Concepts, and Hermeneutical Resistance*

Recent work on hermeneutic injustice emphasizes, contrary to Fricker's original account (2007), that such injustice arises not only from conceptual lacunae in collective interpretive resources,

but also from the presence of impostor concepts that systematically distort the understanding of marginalized groups' social experiences (Deans, 2024; Falbo, 2022; Picazo Jaque & Delgado, 2024). Many instances of hermeneutic injustice in mental health seem to be of the latter type: they involve situations where the hermeneutic tools available to psychiatrized individuals are distorted due to stigma, pernicious stereotypes, or implicit assumptions of medical practice. For example, internalized normalizing or romanticizing attitudes toward depression, which frame it as within the "normal range" of human experience, or as a mark of genius or poetic sensibility, can obscure individuals' proper recognition of their experience and its debilitating consequences (Jackson, 2017). At the same time, the Mad and Neurodiversity movements have long challenged narrow pathologizing narratives that look at divergent cognitive traits exclusively through the lens of deficit (Chapman, 2023; Spandler et al., 2015). This amounts to hermeneutic injustice insofar as it forces individuals to think of their experiences as somehow inherently at odds with mental health and flourishing, blocking alternative neurodiversity-affirming interpretations. In this line, mad activists have also argued that certain depressive traits can be legitimately reframed as "dangerous gifts," a perspective that highlights potential strengths linked to the condition, with potential therapeutic value (Mitchell-Brody, 2007). So, what makes a concept legitimate? How to spot—and whack—moles in our self-understanding repertoire? This talk builds on mindshaping views of self-understanding (Zawidzki, 2016) to address this question. First, against a realist, mindreading-like account of the self-illness distinction, we will argue that distinguishing impostor from legitimate concepts is not a matter of accurately tracking pre-existing self-illness borders (Jeppsson, 2022). Rather, self-understanding is an open-ended and inherently social process, whereby the boundaries between legitimate and impostor concepts are actively constituted in ongoing regulative practices. These involve judging which concepts best fit an agent's values and sense of identity, and how well the agent regulates their behaviour in line with social scripts and normative expectations associated to their use. Instead of an obstacle, we will argue that the absence of definitive borders is a prerequisite for generating new, potentially liberating self-understanding tools. Second, we will argue that this does not implies a "frictionless", "anything goes" view of self-understanding. What self-interpretation tools are at our disposal depends on our social niches, and we cannot change at will their associated scripts and expectations. In this sense, we will argue that a) discerning the legitimacy of a concept requires tracking its genealogy; and b) that legitimate concepts are those crafted and developed in communities of epistemic resistance (Medina, 2013). Engagement with such communities puts agents in a privileged position to ascertain the legitimacy of a concept insofar as it allows them to acquire a form of "double consciousness" (Du Bois, 1903; Toole, 2021): the ability to see the world (and themselves) both through the lenses of hegemonic and counterhegemonic concepts.

Manuel Almagro Holgado & Miguel Núñez de Prado-Gordillo: *Non-ideal Mindshaping & First-Person Authority in Mental Health health.*

Recent trends in analytic philosophy have converged on the need for a non-ideal methodology (Hänel & Müller, 2024). This involves questioning pernicious theoretical idealizations that reflect and reinforce oppressive practices and distort knowledge production—a commitment that, crucially, extends to one's own theoretical perspective (Almagro & Guerra, 2023; Bordonaba-Plou et al., 2022). Here we adopt this self-critical stance toward the mindshaping framework, focusing on background assumptions about first-person authority (FPA)—the deference often granted to individuals' self-attributions of mental states—at its core. Although closely linked to self-knowledge, FPA has long been disentangled from an epistemic reading (McGeer, 1996, 2008; Moran, 2001; Wright, 1998), being instead understood as a social norm governing interpersonal relations (Borgoni, 2025). Specifically, building on McGeer's (1996, 2008) self-regulatory account,

mindshaping theorists usually understand FPA as a default social status conferred upon us due to an agential rather than epistemic privilege: we occupy a privileged position to shape our own minds in line with the social scripts and expectations our self-ascriptions commit us to (Stjernberg, 2025; Strijbos & De Bruin, 2015). This default status, in turn, rests on what these authors, following Wright (1998, p. 632), see as the "telos" of mental interpretation, social coordination: were we unable to systematically rely on each other's ability to "practice what we preach" about our minds, social coordination would collapse. This account of FPA is theoretically compelling—especially when contrasted with traditional epistemic introspective accounts—and politically illuminating, as it highlights the moral dimension of mental interpretation: to paraphrase autism activist Jim Sinclair (1992, p. 302), taking others' self-understandings at face value amounts to granting them the dignity of meeting them on their own terms. Yet the way it's formulated suggests an idealized view of mental interpretation practices, grounded on a dual descriptive-prescriptive idealization: that granting FPA is—or should be—the default attitude toward self-ascriptions. First, this assumption seemingly rests on what Jessica Keiser (2022) calls the "picture of language as cooperative information exchange" (p. 1): a pernicious idealization that treats communication as fundamentally oriented toward social coordination. This picture obscures non-cooperative uses of language, such as hostile or exclusionary speech. In the context of mental interpretation, this means treating exclusionary mental-state attributions, aimed at fostering division or marginalizing certain groups (e.g., attributing beliefs in matriarchy to feminists), as peripheral to interpretive practices—which only makes sense for those privileged enough not to have been systematically subjected to them. One might insist that, even if the idealization is not descriptively accurate, its prescriptive reading is nevertheless necessary to capture what is unjust about undue denials of FPA, like the ones addressed in the other contributions to this symposium. However, we do not think that recognizing this requires universalizing the default status of FPA; instead, some contexts require a default attitude of suspicion, or at least suspension of judgment. We will focus specifically on (a) cases of self-stigma, where harmful self-ascriptions block access to potentially liberating self-understandings, and (b) cases of "therapy-speak," whereby self-ascriptions may function to reinforce oppressive structures (Isern-Mas & Almagro, 2025). Properly handling these cases, we'll argue, calls for a radically context-sensitive approach to FPA and its role in mindshaping.

Thursday 2nd July 14:30 — AI (Grote zolder)

Iwan Williams: *Doing without etiological functions in AI metasemantics*

What, if anything, do the internal states of AI models represent? Do activation patterns in large language models, for instance, represent familiar entities like grapefruits and governments, or merely linguistic objects like words and syntactic structures? Recently, philosophers have turned their attention to these questions, attempting to articulate the conditions under which AI internal states bear particular representational contents. An influential approach (taking cues from teleosemantics) makes appeal to etiological functions – functions deriving from the causal history of a system or lineage of systems (Butlin 2023; Coelho Mollo & Millierè forthcoming; Goldstein & Levinstein 2024). In this paper, I argue that an alternative kind of function – one deriving in part from the intentions of designers, deployers and users – is better suited for the task of grounding representational content in AI systems.

The etiological approach to AI representation draws on theories originally developed for biological cognition (Milikan 1984; Shea 2018). On standard teleosemantic theories, the representational content of an internal state is the worldly condition that the state has the function of carrying information about. These theories standardly employ an etiological notion of function: the function of an internal state derives from the properties it was selected for (e.g. by natural selection), or that were the target of some historical stabilisation process (such as feedback-based learning)

(Garson 2016; Shea 2018). For instance, if activity F in the brains of ancestral frogs contributed to frog proliferation by carrying information about flies, then activity F in present-day frogs has the function of carrying information about flies, and in turn, can be said to represent flies.

Applied to AI, advocates of the etiological approach have argued that machine learning processes (such as supervised learning or reinforcement learning) can constitute the relevant function-conferring selection or stabilisation processes. On this view, internal states that were selected during training for carrying information about model-external features thereby acquire the function of representing those features.

While etiological accounts of function are dominant in theorising about biological functions and biological representation, the invocation of etiological functions in AI metasemantics has not been sufficiently justified. Do the functions that determine representational content in AI systems stem from the etiology of those systems (e.g. their training history) or are they the result of some other function-conferring factor? One reason to question the assumption of the etiological approach is that AI systems are not biological organisms but artifacts, produced, deployed and used by intentional agents for our purposes. So, perhaps human intentions play some role in fixing the (representational) functions of their components.

I identify three desiderata for an account of functions in the context of grounding representational content in AI systems. I then develop an account of intention-derived functions in AI systems, which I dub “deployment functions” and show that they better satisfy these desiderata, compared with etiological functions.

First, I define system-level deployment functions: an AI system has a deployment function to X if agents design, deploy, or use it to X. A chatbot deployed for answering factual questions thus has the (system-level) deployment function of answering factual questions, regardless of whether it was trained on some other task, such as next-token prediction.

Second, I show how system-level deployment functions confer functions on internal components. An initial challenge to developing an intention-based account of functions for components in AI systems is that, unlike ordinary artifacts, the internal workings of deep learning systems emerge through training and are not hand-coded by engineers. Hurshman (2024) thus argues that components of opaque neural networks lack intention-derived functions, since engineers lack the necessary beliefs about component-level mechanisms. However, I argue that intentions can ground component functions indirectly, even when designers are ignorant of component-level mechanisms.

On my account, a component has a component-level deployment function, F, if it contributes to the system’s (system level) deployment function G by F-ing. Crucially, this does not require that designers have any beliefs about the component. Consider ancient Roman concrete, which is unusually durable in seawater. It turns out that – unbeknownst to the ancient Romans – this is due to specific mineral constituents which form fracture-resistant plate-like structures (Jackson et al. 2017). The Romans could not conceive of these micro-structural properties. Yet the minerals, I contend, served the function of forming such structures, because this property is what contributes to the concrete’s intended purpose (resistance to seawater exposure). Similarly, components in opaque neural networks can have deployment functions fixed by their contribution to system-level deployment functions, even when engineers are unaware of what those components do, and independently of any historical facts about the training of the system.

This account of deployment functions meets key desiderata for a theory of representational functions for AI systems: (i) Unlike the related, but weaker notion of a “Cummins function” or “role function” (Cummins 1975; Craver 2001), it exhibits normativity: because the system-level function is genuinely normative (the system is supposed to behave a certain way), components that fail to

make their characteristic contribution to that function are malfunctioning, not merely playing a different causal role. (ii) It accommodates misrepresentation: a component that has the deployment function of carrying information about P, but fails to do so, thereby fails to contribute to the system's deployment function and so misrepresents.

While etiological functions also fulfil the above two desiderata, deployment functions satisfy a third desideratum that etiological functions do not: (iii) explanatory relevance. Interpretability researchers attribute representational contents to AI model-internals in the service of broader pragmatic goals, such as predicting failure modes, mitigating dysfunctional behaviour, and building models that better serve our interests (Sharkey et al. 2025). These concerns are largely ahistorical and impose use-centred success criteria: by positing internal representations, researchers typically want to explain how components contribute to the model's performance on the tasks we currently use them for, rather than the (possibly divergent) roles stabilised through training. Deployment functions are thus more faithful to the explanatory goals and practices of AI interpretability research, and should be preferred over etiological accounts in AI metasemantics.

Dimitri Coelho Mollo: *Representational Explanation for Deep Artificial Neural Networks: prospects and challenges*

The increase in size and capabilities of current deep artificial neural networks (DNNs), and especially of Large Language Models (LLMs), has turned them more and more into explanatory targets in their own right, instead of being seen as 'just' useful tools for assisting in practical and scientific goals. Consequently, the project of AI interpretability, i.e. the scientific field dedicated to explaining and understanding the functioning of AI systems, has grown considerably in the last 5 years or so.

AI interpretability employs a variety of methods to investigate the behaviour and inner workings of AI systems, many of which are inspired or directly copied from methods in cognitive science, such as behavioural testing, ablations, study of stimulus response profiles, and functional modelling. This methodological adoption brings along with it a baggage of conceptual assumptions and explanatory constraints to AI interpretability that have their origin and justification in the practices and aims of cognitive science.

A key, foundational element in mainstream cognitive-scientific methods and explanations is the appeal to internal representations. Such an appeal also permeates AI interpretability work, although its explanatory fruitfulness is debated. In this talk, I examine the prospects for theories of representation for AI systems. I remain neutral on whether such systems should count as cognitive.

Philosophers of cognitive science have dedicated considerable effort to showing that internal representations can be bona fide scientific posits by developing naturalistic theories of representation for biological cognitive systems. Even though there is no consensus on the details of the correct theory of representation, there is wide agreement on its main components. Internal representations are physical states that stand in causal-informational relations to states of the world, and that have the function to carry information about the world, having been selected by natural selection processes to do so.

I identify three main options for theories of internal representation in current AI systems: a) our best theories of internal representation fully apply to current AI systems; b) our best theories of internal representation apply to current AI systems, but require important tweaks; c) theories of

internal representation for current AI systems are substantially different from theories of internal representation for biological systems.

I argue that a) is unlikely to be a promising route, that b) is plausible but the nature, extent, and tenability of the required tweaks remain unclear, and that c) is a potentially philosophically fruitful route, but that it may end up collapsing into b) or leading to a rejection of the use of cognitive science methods in AI interpretability research.

In brief, two main challenges plague strategy a). First, the AI systems to which we apply cognitive-scientific methods today are almost exclusively disembodied pieces of software that lack direct causal-informational relations with the world. The data on which they are trained are, instead, human-produced representations, such as texts and photos. Second, it is unclear that current AI systems undergo the selection processes our best theories of representation require. They do not undergo natural selection, the process that most theories give pride of place to. It is moreover unclear whether DNN training count as genuine learning, or at least the kind of learning that can ground representational functions.

When it comes to b), it seems that at least two key tweaks are required. A theory of internal representation for current DNNs should not require direct causal-informational relations to the world, but instead show that exclusively indirect, human-mediated causal-informational relations are sufficient. This is far from a trivial endeavour, as it needs to be shown that AI representations can be about the world despite this mediation, rather than being 'just' representations of human representations. In addition, the theory would need to appeal to non-natural selection processes, perhaps learning-like processes during training. However, an analogous difficulty appears here: even conceding that appropriate selection processes are at play, it must also be shown that the selection 'rationales' are worldly states, rather than purely human representations of such states. Should strategy b) fail to meet these twin challenges, representational explanation for AI systems would at best be fundamentally different to cognitive science explanations: AI representations would be exclusively representations of certain human populations' representational space, rather than representations of the world – with consequent differences in our understanding of AI systems and of whether they can count as cognitive.

Finally, strategy c) is partly motivated by the difficulties pointed out above: the differences between biological systems and DNNs are such that we may need fundamentally different theories of representation for AI systems. On this view, DNNs are a 'new kind of beast', sharing some core features with biological cognitive systems, while lacking others. It is unclear what factors could constitute the grounds for such alternative, DNN-focused theories of representation. If they also end up appealing to causal-informational relations and selection processes, the strategy risks collapsing into b). If, on the other hand, such theories are indeed fundamentally different to theories of representation for biological systems, they run the risk of falling outside the purview of the explanatory practices and methods of cognitive science, thus putting pressure on the extension of cognitive-scientific methods to the study of AI systems. That said, the proof of the pudding is in the eating. Philosophers of AI should try and develop candidate DNN-focused theories of representation, so that we can assess their tenability and independent scientific fruitfulness.

In this talk my job has been mostly that of presenting a space of possibilities and their associated challenges, but it is also a plea for trying out new, unbeaten paths in our philosophical exploration of representation and cognition over and beyond biological systems.

Tom-Felix Berger: *The Conceptual Role of Internal Representations in Strategic AI Deception*

Human deception involves intentionally causing false beliefs. Since it is contested whether large language models (LLMs) and artificial intelligence (AI) in general have beliefs and intentions, deception by such systems is typically defined in terms of observable behavior. However, this paper argues that such behavioral definitions fall short of the strategic nature of deception, among other things, because they do not consistently distinguish deception from mere error. Moreover, behavioral definitions of AI deception are not optimal as methodological guidance for mechanistic approaches to AI deception because they do not specify candidates for internal markers of deception. It is important to take the internal mechanisms of LLMs and AI systems into account when addressing deception, since skilled forms of deception may be behaviorally indistinguishable from honesty.

This paper formulates these challenges to behavioral definitions and addresses them by proposing a novel definition of strategic AI deception in terms of generalized propositional attitudes. Specifically, it provides generalizations of the attitudes of belief and desire that 1) can be ascribed to current and future AI systems less controversially than their human counterparts, 2) are suited to characterize the subtleties of strategic deception, 3) are relevant to large-scale risks of AI deception, and 4) are methodologically instructive for mechanistic approaches to deception.

Maja Sostaric: *Basic emotions and AI*

I contend that the widespread use of Paul Ekman's theory of basic emotions in artificial intelligence is scientifically and philosophically questionable, despite its apparent practical utility. In this paper, I briefly outline the key assumptions of Ekman's theory and selected critiques of emotional essentialism as contextual background, before focusing primarily on whether the revised position preserves any explanatory or empirical value in light of sustained criticism.

Researchers still cannot agree on a specific definition of emotion, but many classifications of this psychological phenomenon have been developed. One of them, also known as the "basic emotion theory", assumes that: There are certain universal basic emotions in every human being These can be identified in an individual through the analysis of facial expressions and microexpressions.

The author of this concept is Paul Ekman, whose research findings, despite criticism from anthropologists, psychologists, neuroscientists, and philosophers, have become firmly anchored in much of contemporary reality: national security systems at airports, education, hiring start-ups, systems that purport to detect psychiatric illness and policing programmes that claim to predict violence. Existing research indicates that there are flaws in his model, which they assume to be overly simplistic, as many reactions and descriptions of emotional experiences do not fit into these categories (Coppini et al., 2023). Furthermore, Barrett et al. (2019) argue that there is no consistent evidence for a relationship between facial configuration and a specific emotional state, and that context, experience, and the observer's perceptual abilities influence the interpretation of a given emotion.

Based on Ekman's research, emotion recognition was developed, which aims to identify and interpret human emotions and mental states by analysing various cues, such as facial expressions, body language, and speech patterns.

Nowadays it is mostly a subfield of AI that by leveraging machine learning algorithms and computer vision techniques, it attempts to classify emotions into categories such as happiness, sadness, anger, fear, surprise, and disgust (Crawford, 2021). It can be found in programs and technologies like Affectiva - a software company that has developed an emotion recognition

platform using computer vision and deep learning algorithms to analyse facial expressions in real-time, with applications in for example marketing, automotive, and gaming. In marketing, companies in the US, Europe, and Asia use Affectiva's Emotion AI to analyse viewers' reactions to advertisements, films, and TV programs. The system evaluates viewers' facial expressions and emotions to optimise content and the effectiveness of advertising campaigns. Another program is SPOT (Screening of Passengers by Observation Techniques), which detects threats at airports by analysing passengers' behavior and microexpressions. Other AI recognition programs can be found in recruitment, by companies such as Intel, which during job interviews try to detect a certain emotion based on facial expressions, then they link the assumed emotion with temperament characteristics, and later with the competencies that the company is looking for in an employee.

But why has the idea that there is a small set of universal emotions, easily interpreted from facial expressions, become so widely accepted in the field of artificial intelligence, despite numerous pieces of evidence to the contrary? It is said that it offered a small set of principles that could be applied everywhere, a simplification of complexity that was easily replicable and it can help in influencing behaviour in training people to perform in recognisable ways.

In my presentation, I begin with current contemporary applications based on Ekman's work. Later, I show that behind this there are specific assumptions that can be called emotional essentialism, which assumes that different emotions have their universal, ideal essences, understandable to every human being. Next, I move on to critical aspects of essentialist theories, with particular emphasis on Ekman's theory and, consequently, industries that use AI. I will focus on the criticism of researchers such as social scientists Maria Gendron and Lisa Feldman Barrett, and AI expert Kate Crawford. I conclude my paper with the most important question and possible answers: Can any value be found in Ekman's concept, despite the criticism? Ekman has revised his original claims over time. In this paper, I address the motivations underlying these revisions and assess the extent to which they justify the claim that his theory is consistent with experimental data. I argue that this is not the case, which supports competing paradigms in emotion theory, in particular the constructivist approach developed, among others, by Lisa Feldman Barrett.

Thursday 2nd July 14:30 — Memory (Spiegelzaal)

José Carlos Camillo: *Assessing the reliability of memory-based beliefs of religious experiences*

In a recent paper, Munro (2024) addresses two questions concerning religious beliefs that originate in religious experiences, understood as experiences whose objects are supernatural entities or events. The first is the Psychological Question (PQ), which concerns the causal relationship between religious experiences and religious beliefs. The second is the Reliability Question (RQ), which asks whether this psychological role is reliable, thereby yielding reliably formed religious beliefs.

The most common response to PQ holds that religious experiences play a role analogous to that of perceptual experiences in the formation of perceptual beliefs. Munro criticizes the perceptual answer to PQ on the grounds that it makes RQ dependent on much more difficult questions. Assessing the reliability of religious experiences would require determining whether they are perceptual or hallucinatory, which depends on whether supernatural entities/events exist and can be perceived. Moreover, if perception and hallucination are the same type of process, they would have the same reliability. These are difficult questions a perceptual account of PQ must answer before addressing RQ. Munro proposes a memory-and-reflection account of PQ, on which RQ

can be addressed more easily, since the reliability of memory and reflection can be evaluated independently of these difficult issues.

In this talk, I pursue two aims. Negatively, I argue that Munro's attempt to avoid these "difficult questions" either returns us to them or depends on different, but still difficult debates. Positively, I suggest that the reliability of religious beliefs formed by remembering religious experiences should be assessed on a case-by-case basis, since remembering different kinds of religious experiences can be more or less prone to error, and remembering religious experiences can reliably support some kinds of beliefs but not others. The reliability of a belief depends not only on the reliability of the final process that produces it, but on the reliability of the entire causal chain leading to it. Even a valid inference, which is a reliable process, can result in an unreliable belief if one of its premises originates in an unreliable process, such as wishful thinking (Goldman, 2021). For this reason, Goldman (1986) and others have argued that the reliability of memory beliefs depends on the reliability of the experience that originally caused the memory, given a preservationist view of memory. On this view, assessing the reliability of memory does not help answer RQ, since memory-based beliefs inherit the reliability of the experience. We are thus pushed back to the "difficult questions" he aimed to avoid. Recent developments in the philosophy of memory might seem to offer Munro a way to avoid this result. He could, for instance, appeal to post-causalist views, according to which memory need not be caused by the remembered experience (Michaelian, 2024), or to generativist views in the epistemology of memory, according to which memory generates new justification independently of perception (Tooming & Miyazono, 2024).

However, adopting these views merely replaces one set of difficult questions with others, namely those raised in the (anti-)causalism debate and the generativism/preservationism debate. Moreover, Munro would also need to take a stand on the debate about the accuracy of memory: whether memory aims to accurately represent the past event (alethism), the past experience (radical authenticism), or both (authenticism) (McCarroll & Perrin, 2025). This choice matters because it determines which kinds of beliefs can be reliably formed. Radical authenticism licenses reliability only for beliefs about past experiences (e.g., "I had a religious experience"), whereas alethism licenses reliability only for beliefs about the world (e.g., "There are supernatural entities or events"). Thus, assessments of reliability vary significantly depending on the view of memory accuracy adopted. For these reasons, Munro's proposal that the reliability of religious beliefs can be assessed by evaluating the reliability of memory processes either returns to the same questions or depends on answering other difficult questions. It is therefore unclear whether his strategy genuinely makes RQ more easily answerable. That said, there is an alternative route. Empirical studies on the accuracy of remembering religious experiences can be used to assess the reliability of beliefs about past experiences without settling foundational questions about the nature of memory. Such studies show, for example, that factors like emotional arousal and emotional regulation affect how accurately religious experiences are remembered (van Mulukom, 2017). This suggests that different emotional processes have distinct epistemic consequences for belief formation. Accordingly, the reliability of particular religious beliefs should be assessed in light of the specific processes that influenced memory formation in each case, in particular, emotional processes.

Sacha Behrend: *Reference and Representational Formats in Episodic Memory*

On simulationist accounts of episodic memory, remembering a past event does not require the preservation of an appropriate causal connection between the original experience and its subsequent retrieval. Instead, episodic remembering is understood as the product of a reliable episodic construction system that simulates experiential episodes belonging to the subject's

personal past. While this framework offers a powerful alternative to causal theories of memory, it faces an explanatory challenge: if episodic memories are not causally connected to past experiences, how can a given memory be about one particular past event rather than another event of the same type?

Openshaw and Michaelian (2024) address this challenge by proposing a broadly reliabilist model of mnemic reference-fixing. Drawing on work on singular thought, their account explains reference not in terms of causal chains, but in terms of reliable, non-accidental relations between mnemonic representations and features of past events.

In this paper, I argue that this model overlooks a crucial dimension of episodic memory, namely the diversity of its representational formats. Both philosophical theorizing and empirical research leave open the possibility that episodic memories do not all share the same representational format. In addition to propositional representations, episodic memory has also been modeled as involving depictive or imagistic representations, among others. Because representational format constrains the ways in which a mental state can refer to events and properties, the question of format is directly relevant to evaluating the scope and adequacy of any model of mnemic reference-fixing.

I distinguish at least three candidate formats for episodic memory—propositional, depictive, and analog—and examine whether Openshaw and Michaelian's model, and in particular what they call the referentiality and veridicality conditions, applies equally well to each. I argue that the referentiality condition is well suited to propositional episodic memories, which explicitly attribute properties to past events and are therefore naturally evaluated in terms of truth and error. By contrast, difficulties arise in non-propositional cases. Depictive and analog episodic memories do not explicitly, or even implicitly, attribute all relevant properties. Instead, they also rely on what I call interpretive properties, which are not directly encoded in the representational format itself.

I show that the original referentiality condition does not adequately capture how such interpretive properties are fixed. While depictive representations can explicitly or implicitly encode spatial and visual features, and analog representations can covary with certain magnitudes, neither format by itself allows to attribute emotional or conceptual properties. To account for these cases, I propose an amendment of the referentiality condition that distinguishes between attributable properties, which are fixed through explicit or implicit attribution by the episodic construction system, and interpretive properties, which are fixed through reliable interpretive processes operating on the output of that system.

I then turn to the veridicality condition, which is meant to account for how a successful episodic memory accurately represents the relevant past event. I argue that this condition also requires amendment once the diversity of representational formats is taken seriously. In particular, while propositional episodic memories can be evaluated in terms of truth and falsity, depictive and analog memories are more naturally assessed in terms of degrees of accuracy. Moreover, different formats impose different constraints on what counts as accurate remembering, since not all formats are capable of representing the same properties. I therefore propose revised versions of the veridicality condition that (i) allow for graded accuracy, (ii) reflect the dependence of accuracy on representational format, and (iii) distinguish between accuracy with respect to attributable properties and accuracy with respect to interpretive properties.

The resulting account preserves the core simulationist–reliabilist insight that reference in episodic memory is fixed by reliable cognitive processes rather than by causal traces, while

accommodating the heterogeneity of mnemonic representation. I conclude by considering and responding to worries about this proposal.

Matheus Diesel Werberich: *Why do we have two memory systems? An information-theoretic account*

Traditional taxonomies of long-term memory mark a distinction between episodic memory, the capacity to relive events in one's personal past, and semantic memory, the capacity to remember general facts about oneself or the environment. Recently, such distinction has fallen under heavy scrutiny (Aronowitz 2022; Gentry & Buckner 2024). Given the interactions between episodic and semantic memory under consolidation, and the fact that some memories present both episodic and semantic features, philosophers and cognitive scientists have argued that there is no clean distinction between these memory systems. Instead, these researchers argue that episodic and semantic memory form a continuum. However, what is missing from this literature is an account of *why* this continuum exists and of the computational principles that underlie it. In this paper, I defend that such continuum is adaptive due to the constraints of rate-distortion theory.

Rate-distortion theory is a formal account of the idea that for any representation to be as faithful as possible, it must be as complex as its target. For example, a map of a city will be more accurate the more it encodes details about the city itself. Yet, doing so would diminish its usefulness as a map. As such, there is a trade-off between representational complexity and fidelity. In line with this trade-off, rate-distortion theory asks how many bits of information can one remove without distorting a source beyond a certain point. Where $R \in \mathbbm{R}$ is a representation's rate (roughly equivalent to its length and complexity), and $d : X \times Y \rightarrow \mathbbm{R}_{\geq 0}$ is a function that measures how much distortion there is between a representation and its target, Shannon (1959) proves that, for any distortion value $d^\prime$, there is a single lowest value of R that meets the distortion $d^\prime$, denoted as $R(d^\prime)$. Likewise, for a set rate R, $d^\prime$ is the minimum amount of distortion one can expect for a representation of that complexity.

Rate-distortion theory thus places a hard constraint on information processors. It stipulates that, for any distortion between a representation and its target, the complexity of that representation must be at least $R(d^\prime)$. Like other constraints, it sets hard limits on the value of representational complexity. Moreover, the theory is external to how an information processor is implemented, is fixed relative to other factors in the vicinity and guides the explanandum outcome, rather than triggering it. Per Ross (2023), these features are characteristic of hard constraints, which explain why it is impossible for some systems to feature some properties or instantiate some values.

Being systems that process information from the past, episodic and semantic memory have to follow the constraints of rate-distortion theory. Moreover, the continuum between these systems naturally mirrors the range of possible pairs of rate and distortion values. Whereas episodic memories are complex representations (i.e., with higher rates) that faithfully reconstruct the past (hence, with a lower distortion), semantic memories are simple representations (lower rates) that do not capture the details of past events (higher distortion). Having both of these memory systems is adaptive in so far as they capture both ends of the trade-off: they allow for us to use simple representations when accuracy is not paramount, while having the option to use a more complex representation when we need less distortion. It also allows us to make generalizations and learn patterns without the cost of losing specific information about past events.

Moreover, one consequence of rate-distortion theory is that overlapping representations are more efficiently encoded as one representation (Berger, 1971). Some experiments suggest that our

memory systems function in this way. The Mnenomic Similarity Test was designed specifically to measure participants' ability in recognizing old stimuli and telling them apart from similar, though different, items (Kirwan & Stark 2007). Different studies in this paradigm have consistently shown that target repetition in the encoding phase is correlated with participants' misclassification of lures as old items (Reagh & Yassa 2014). This is predicted from rate-distortion theory: correlated and overlapping messages can be lumped into a single representation for efficient encoding. While this happens at the cost of higher distortion, it is a way for our memory system to spare the amount of resources dedicated to remembering the past.

In summary, rate-distortion theory provides a formal account of why episodic and semantic memory systems form a continuum. According to this account, such continuum is adaptive because it allows for our cognitive system to flexibly use representations that fall on either end of the rate-distortion curve.

Albert Newen and Denis Perrin: *What does it involve to re-remember an event? Functionalizing re-remembering*

Our memories not only occur, they also reoccur: we re-remember. What is the standard according to which an episodic recall can be classified as a re-remembering of a former episodic memory? We argue that standards of re-remembering depend on the functional roles of episodic memory with the well-established functional trias of enabling improved orientation in the world, of fostering social understanding and of keeping and enhancing a positive self. Focusing on these three roles allows us to establish the main claim: Whether a case of episodic recall is a case of re-remembering an event depends on the explanatory interest which is mainly depending on the contextually indicated dominant functional role of the recall in a situation. To work this out we describe prototypical situations in which a recall can be embedded, e.g. a legal situation, a social situation of shared narratives and a personal situation of transformative experiences. These situations are used to develop several standards of genuine re-remembering which can lead to different evaluations of the same type of recall. Thus, we on the basis of a functionalizing perspective of episodic memory we propose a context-dependent account of re-remembering such that a situation typically determines a dominant functional role and thereby the relevant explanatory interest connected to the episodic recall. To develop this in detail, we propose a constructive account of episodic memory with the core claim that episodic memory is the product of scenario construction which starts with a sensory input which triggers a memory trace. Then the memory trace is enrichment by semantic information (semantic memory) to result in a full vivid recall of a scenario. And re-remembering an episode is relying on the same process again.

Figure_1: Episodic re-remembering. (see PDF)

Furthermore, we propose to distinguish four dimensions of the recall of an episodic event and typical implementing features of each event with the aim to develop standards of evaluating whether a recall is a case of re-remembering a past event or not. We propose to distinguish the four dimensions, namely the representational vehicle, the representational format, the situational content and the experiential content. Each of these dimensions has a variety of implementing features. Figure 2. Characterization of the episodic recall of an event with its dimensions and their implementing features (see PDF) On this basis we describe three standards of re-remembering. In line with the literature, we presuppose that there are three functional roles of episodic recall, namely the directive function to foster adequate acting in the world and planning future action, the

social function to improve the understanding and prediction of others, and the self-function to keep and enhance a positive and coherent self-understanding (Bluck et al., 2005).

On this basis, we can outline three corresponding prototypical contexts of recall and develop them into three standards of re-remembering illustrated by typical contexts: (i) The legal context of a testimony aiming at veridically describing the world: A witness is supposed to deliver veridical information about a past event, e.g. a car accident, anchored in the experience of the event for the sake of the first-handedness and reliability of information. We are then interested in a veridical description of the actual event and its components from an ideally objective perspective.

(1) The standard of informative veridicality: An episodic recall EM2 (at t2) is a case of re-remembering concerning informative veridicality with respect to an episodic recall EM1 iff: (i) both EM2 and EM1 are based on the same neural vehicle which is anchored in the relevant past event e; (ii) EM2 and EM1 involve representations of a similar type of situation including objects and properties to a sufficient degree of similarity; (iii) these representations overlap sufficiently with the actual components of e. (ii) The social context of a shared memory-based narrative aiming at understanding others: Two colleagues have participated in a faculty meeting and agree afterwards upon a description of the relevant aspects of their episodic memories of the event. These shared narratives can be understood as sharing their situational content and it is a good indicator of their shared cognitive and behavioural dispositions concerning the relevant aspects of the event. This makes their behaviour predictable and understandable for each other.

(2) The standard of intersubjectively shared narratives: An episodic recall EM2 is a case of re-remembering concerning intersubjectively shared narratives with respect to an episodic recall EM1 iff: (i) both EM2 and EM1 are available in a narrative format of representation; (ii) the narrative contents of EM1 and EM2 are sufficiently overlapping as far as the situational content is concerned; (iii) ideally their content are also sufficiently overlapping in experiential content as far as the fine-grained conceptualization expressed in the shared narrative is concerned. (iii) The personal context of aiming at keeping or enhancing a (normally) positive self-image: A father remembers the first family holiday on the Atlantic coast which became a regular, annual holiday at the same campsite for ten years. Due to the repeated visits, re-remembering this first holiday may involve many mistakes due to the integration of objects, properties etc. from later iterations, while a minimal reference remains to the relevant spatiotemporal unit, the campsite on the first holiday. What matters to the father is only that all his recalls are integrated with this unit, and most importantly, that the same experiential features are activated, like the richness of the bodily, emotional and agential experience, together with a sense of the high level of relevance for his self-identity. (3) The standard of experiential constancy: An episodic recall EM2 (at t2) is a case of re-remembering with experiential constancy with respect to an episodic recall EM1 iff (i) EM2 and EM1 involve representations of minimally similar situational content, typically overlapping spatiotemporal units, and most importantly (ii) EM2 and EM1 share their experiential features to a high degree.

To conclude: we have argued that whether a given episodic recall qualifies as a case of re-remembering depends on the contextually-relevant interest in a given situation and the relevant corresponding functional role activated by this interest.

Thursday 2nd July 14:30 — Pragmatics (Voorkamer)

Maria Spychalska, Ludmila Reimer and Markus Werning: *Quantifier and implicature processing in contexts with full and partial information*

See PDF.

Maria Spychalska: *Implicatures in Epistemically Uncertain Contexts: What Does the Evidence Tell Us?*

See PDF.

Christian De Leon: *In the Zeitgeist: Common Ground and Collective Attention*

Ordinarily, felicitous use of an anaphoric phrase requires that there be a linguistically introduced antecedent. However there exists a class of cases that violate this rule of thumb, for example (1) might be uttered discourse-initially: (1) There's no way he'd declare war on Greenland. Right?

To resolve the referent of "he", the hearer must be aware of recent news and take the speaker to be similarly aware. Standard theories suggest these events must be *common ground* (Stalnaker 1978, 2002, 2014; Clark 1996), requiring conversational participants to mutually know, believe, or accept certain facts. However, these accounts face a coordination problem: in cases like (1), or social media posts like (2), speakers often lack the evidence necessary to justify an iterative mutual propositional attitude.

(2) Sooo do we think jumbotrons are gonna be things of the past now?

In social media contexts, this is exacerbated by *context collapse* (Marwick & boyd 2011), where the speaker cannot verify the audience's knowledge. Furthermore, for such utterances to succeed, the event must not only be known but achieve a high degree of *mutual manifestness* (Sperber & Wilson 1986) or *salience*. Out of context, (2) is difficult to interpret. But when viewed in July 2025, when a CEO went viral for being caught on the kiss cam at a Coldplay concert apparently having an affair with another company executive, (2) makes sense. In February 2026, even if one were aware of last summer's incident, it would be considerably more effortful to interpret the message. Unlike in e.g. cases in which a goat walks into the room in which a conversation is taking place (Stalnaker 1978, p.86), it is difficult to see how the interlocutors in (1) and (2) could achieve *grounding* (Clark & Brennan 1991). In cases of social media communication such as (2), the speaker does not even know who the audience is or if there is one at all.

In this paper, I argue that the requirement of standard views that interlocutors hold some kind of mutual propositional attitude is too strict. I offer an account designed to make sense of cases like (1) and (2) which prioritizes states of *collective attention*. Drawing on recent trends in cognitive psychology (Brennan et al. 2010; León 2021; Shteynberg 2015, 2018; Shteynberg et. al. 2020, 2023; Tollefsen & Dale 2012), I argue that communication of the kind operative in (1) and (2) depends on cognitive states which represent interlocutors as co-attending to some event. Such attentional states crucially involve first person plural representations—they are not representations of "you" and "I" co-attending, but of "us" as attending. On the account I develop, in uttering (1) or (2), the speaker can satisfy the requisite attentional state---attending to the relevant event via a plural "we" representation---without possessing the kind of evidence that would be required to justify a mutual propositional attitude such as knowledge, belief, or

acceptance. On the hearer side, when uptake is achieved, they come to attend to the event in a first person plural mode as a result of hearing/reading the utterance.

I compare my view with a recent proposal by Lewis (2025), which is designed to update standard common ground theories in light of cases like these. On her view, communication goes by way of an *imagined common ground*---instead of actually achieving a state of mutual knowledge/belief/acceptance, the speaker merely imagines that such a state has been achieved.. I show that while our proposals are compatible, the unique advantage that my proposal enjoys is that it can provide an explanation of the cognitive transition a speaker undergoes when they decide to deploy an anaphoric phrase discourse-initially: they shift from a first person singular attentional sate ("I") to a first person plural one ("we"). Whereas an imagined common ground approach continues to require on overt representation of an (imagined) audience, the collective attention approach analyzes speakers as attending to relevant events in distinctly plural modes (Gallotti & Firth 2013; Schmid 2014.

Alan Lombardini, Giulia Giunta, Diana Mazzarella and Didier Maillat: *At-Issueness in Misinformation Detection: An Investigation of Disfluency in L2 Processing"*

This preregistered study examined the effect of cognitive load stemming from second language (L2) processing on misinformation detection as a function of its at issueness. Previous work showed that false information is detected faster and more accurately when at issue (Giunta et al. 2025). We extended this line of research to L2 processing, exploring how heightened cognitive load influences misinformation detection. As L2 is typically more effortful to process (Foster-Cohen 2000, Padilla Cruz 2013), the resulting cognitive strain may alter how cognitive resources for misinformation detection are allocated. The identification and processing of at-issue content being paramount, misinformation detection in L2 may rely on a special optimization procedure that strategically reallocates limited cognitive resources toward at-issue information, thereby prioritizing the scrutiny of more relevant content (Sperber et al. 2010). We call this the optimization hypothesis. Alternatively, difficulty integrating syntactic information in L2 (Sorace 2011) may impair the identification of at issue content, as syntactic structure encodes propositional prominence (Gutzmann 2023). On this view, misinformation detection in L2 may be less sensitive to the distinction between at-issue and not-at-issue information. This would result in a more uniform level of scrutiny of at-issue and not-at-issue information, thereby reducing the impact of linguistic framing on misinformation detection. We refer to this view as the deprecation hypothesis.

To test these hypotheses, we adapted a misinformation detection task from Giunta et al. (2025) in a 2 (Nativeness: native vs. non native) × 2 (At issueness: at issue vs. not at issue) mixed design. Four hundred participants recruited on Prolific first read a brief crime report and then evaluated a dialogue between a policeman and an informant. Their task was to judge whether the informant's answer was true. Twelve items were presented: six controls and six critical items in which the informant provided false information. At issueness was operationalized by structuring the policeman's question so that it syntactically selected the false proposition as the Question Under Discussion (QUD) (the at-issue condition) or rendered it non-negotiated and backgrounded (the not-at-issue condition). Control items were a version of the informant's answer with no erroneous information. We measured Accuracy and Response Times of misinformation detection. Nonnative participants (N=200) were asked to take the LexTALE language competence test (Lemhöfer & Broersma 2012) upon completion of the main experiment tasks. We expected to replicate the results of Giunta et al. (2025), showing that in L1 at-issue misinformation is detected more accurately and faster than not-at-issue information. For misinformation in L2, we investigated the predictions of the optimization and deprecation hypotheses. On the one hand, the optimization

hypothesis predicts an interaction between At-issueness and Nativeness, with higher accuracy and faster responses for at issue content in L2 than in L1.

On the other hand, the deprecation hypothesis posits that increased processing effort in L2 blurs the distinction between at issue and not at issue material, thus predicting an interaction between At-issueness and Nativeness, with a smaller asymmetry in detection rates and response times across conditions in L2 than in L1. Accuracy results using Generalized Mixed Effects Models in R showed a major drop in not-at-issue content ($p < .001$). In addition to this, non-natives outperformed natives in misinformation detection in both conditions ($p = .029$). About Response times, statistical tests on correct responses showed that all participants were slower in the not-at-issue condition than in the at-issue condition ($p < .001$) and that non-natives were slower overall ($p = .0018$). A post-hoc correlation model between accuracy and LexTALE scores showed that higher competence leads to higher accuracy overall ($p = .009$). The model showed an interaction between At-issueness and LexTALE scores ($p = .038$), whereby competence correlates positively with accuracy in at-issue trials, even though a second model between response times and LexTALE scores showed that higher language competence leads to faster RTs ($p < .001$). Non native participants' LexTALE scores ranged from 50 to 100. Using a cut-off point of 70, we divided them into Low and High competence groups and compared these groups with Native speakers in a simplified correlation model between accuracy and LexTALE scores. The model showed main negative effects for the Native ($p = .0013$) and Low ($p = .0013$) groups, against the baseline of the High group which had the highest accuracy. Estimated marginal means further revealed that whereas High and Native participants differentiated clearly between conditions ($p < .0001$), the Low group showed no significant contrast, suggesting reduced sensitivity to At issueness and aligning with the deprecation hypothesis. However, pairwise comparisons in the at issue condition showed significant contrasts between the High group and both the Native ($p = .0038$) and Low ($p = .0039$) groups, whereas the latter two did not differ.

This pattern indicates that high competence learners showed the highest accuracy on at issue trials, consistent with the optimization hypothesis. At-issue content consistently facilitated misinformation detection across the L1 and L2 groups, with higher accuracy despite faster responses in at-issue trials. L2 participants were slightly more accurate overall, aligning with evidence that cognitive strain can enhance analytical thinking (Alter et al., 2007). Indeed, according to Alter et al. (2007, 569), "[m]etacognitive experiences of difficulty or disfluency appear to serve as an alarm that activates analytic forms of reasoning that assess and sometimes correct the output of more intuitive forms of reasoning." The results of our correlation models reveal a graded effect of L2 competence on at-issue sensitivity, rather than strict support for either the optimization or deprecation hypothesis. As cognitive load reduces as a result of increased language competence, information processing shifts from a non-discriminatory approach to an optimized treatment targeting relevant information.

Thursday 2nd July 14:30 — Valence & Desire (Bovenkamer)

Krzysztof Dolega: *Is valence a unitary and amodal natural kind?*

Valence is commonly understood as the positive or negative character of emotions, affective states, and, according to some accounts, even perceptual experiences. Although it is considered to be one of the central concepts in affective sciences and philosophy of emotion, there is widespread disagreement about its underlying nature (Colombetti 2005). Philosophers debate whether differences in valence stem from differences in mental states' contents (Martínez, 2011; Bain, 2013; Klein, 2015; Carruthers, 2018), the attitudes directed at those contents (Jacobson,

2021; De Vignemont, 2023), or the functional roles they play in the wider cognitive economy (Aydede & Fulkerson, 2018).

Despite these disagreements, two assumptions are widely shared in the literature. First, valence is taken to be unitary, i.e. it is identified with a single psychological kind common across affective states. Second, it is assumed to be amodal, i.e. “nonsensory, or not specific to any perceptual modality" (Carruthers, 2018, p. 668). Even theorists who allow that perceptual experiences can be valenced typically endorse this second assumption, since "Valence does not look like anything, or it looks like too many things" (De Vignemont, 2023, p. 10).

This paper challenges both assumptions by drawing on empirical studies from neuroscience and psychology. First, I argue that the processes underpinning valence are not exclusively amodal but rely on both modality-specific and modality-general neural encodings. Evidence for modality-specific valence processing comes from multiple sensory domains (Miskovic & Anderson, 2018). For instance, neurons in the olfactory epithelium exhibit a patchy organization with a valenced tuning profile (Lapid et al., 2011), while the majority of neural responses associated with gustatory valence is not shared with those correlated with visual valence (Chikazoe et al., 2014). Recent work further suggests that even within vision, valence processing depends on stimulus type: although it is possible to find differential responses to valence of negative and positive stimuli in early visual cortices, these neural signatures are highly dependent on the nature of the stimulus itself (Ballotta et al., 2023). This picture is reinforced by behavioral evidence showing Garner interference between hue and valence (Jacobson et al., 2024), indicating that perceptual valence is not processed independently of modality-specific sensory features.

I then argue that modality-specific and modality-general valence processing correspond to distinct natural kinds. In particular, I suggest that the widespread belief in the unitary nature of valence results from conflating two different phenomena: the valenced character of affect-laden experiences and domain-general evaluative processes involved in decision-making. This hypothesis is supported by the fact that neural evidence for modality-general “valence” substantially overlaps with evidence for abstract value representations used to compare and trade off disparate kinds of options (De Martino & Cortese, 2023). Crucially, both kinds of processes reliably recruit overlapping brain regions, most notably the ventromedial prefrontal cortex and the orbitofrontal cortex (Bartra, McGuire, & Kable, 2013; Sescousse et al., 2014). I argue that what has been described as amodal valence is better understood as value: an abstract representational format functioning as a neural “common currency” that enables comparison between otherwise incommensurable options (Levy & Glimcher 2012). Valence and value, on this view, are not different realizations of a single kind, but two distinct psychological kinds that have been systematically conflated. I conclude by addressing several objections to this hypothesis.

Pietro Chiericoni: *How Emotionally Valenced Properties Inform Perceptual Content*

Whether we are capable of perceiving higher-level properties or not is one of the leading debates in the philosophy of perception [Siegel 2006, 2010]. It is in turn controversial whether we possess the ability to represent emotionally charged properties, i.e., properties that attribute emotional valence to objects. The aim of this paper is to argue that we are indeed capable of representing emotional properties at the perceptual level. Specifically, I will argue for this conclusion based on recent developments both in (i) neuroscience and (ii) cultural neuroscience. These studies are meant to support the thesis by showing that the brain regions involved in perceptual processing show heightened activity when presented with emotionally valenced objects. Seeing a man with a balaclava running towards you elicits a stronger cerebral reaction compared to seeing a robin flying. This heightened activity is not only detected in the regions that are normally associated with emotional processing - such as the amygdala, the hypothalamus, the limbic system and some

of its related subcortical areas - but can also be observed in a variety of large-scale networks of brain regions; some already known to contribute to emotional responses, and others not previously associated with this role. Similar insights have emerged regarding the relationship between sensory and emotional processing, suggesting that these domains are far more integrated than once believed [Murphy 1956, Shuler and Bear 2006, Pessoa and Adolphs 2010, Lebrecht et al. 2012, Chikazoe et al. 2014, Mentec, Ivanchei and Cleeremans forthcoming]. As far as cultural studies are concerned, this paper will examine a wide range of empirical studies on perception [Derntl et al. 2012, Chiao et al. 2016, Iidaka and Harada 2016, Clobert and Tsai 2019] which will be used to support the main claim of the paper. In this respect, cross-cultural studies provide a contrastive line of support for the thesis. They show that individuals from different cultural backgrounds process and, as a result, represent the same emotional stimuli in different ways. Regardless of whether these stimuli are ultimately represented as bearing a negative or positive emotional valence, they elicit remarkable neural responses in sensory cortices, once more confirming that emotional properties are already represented at the perceptual level.

It could be reasonably argued that the enhanced activity detected in sensory areas of the brain can be fully accounted for by patterns of low-level perceptual properties, without requiring any appeal to higher-level ones. Returning to the previous example - the shape, light, speed and colors of the presumed assailant would be enough to elicit a distinctive response in our perceptual system; accordingly, appealing to higher-level properties would be, at best, unnecessary and, at worst, misleading. While this position has some intuitive appeal, I argue that low-level properties alone cannot adequately account for such cases. Drawing on Kragel [2019], I propose two philosophical interpretations for cases like the one proposed. According to the first, it could be claimed that when we perceive objects that strike us, the brain regions deputed for sensory processing get triggered in a stronger manner. This sense of engagement consequently triggers other areas - those specialized for emotional processing - which actually form the emotionally valenced representation. In this case, emotionally valenced properties are primarily represented at the cognitive level, but top-down influences on lower-level perceptual states still inform us on the affective valence of the perceived object.

On the second interpretation, emotionally valenced representations are directly formed within the perceptual system. When presented with an emotionally salient object, we immediately represent it as such instead of elaborating it at the cognitive level. I argue that these accounts should not be treated as competing explanations but as complementary components of a single framework. Through repeated top-down modulation, perceptual systems can become increasingly sensitive to emotional properties, such that objects not initially recognized as emotionally valenced can, through experience, come to be directly represented as such at the perceptual level [see for example Stokes 2021]. When the empirical data are interpreted in this light, the most plausible explanation of the observed patterns is that emotional properties are attributed during perceptual processing itself.

My argument will take the form of an inference to the best explanation. In conclusion, this paper defends the thesis that emotional properties can be perceptually represented. The variety of studies addressed here seem to ground such claim. They also seem to point to a more general conclusion. If empirical evidence suggests that we are capable of representing emotional properties, and if emotional properties constitute a subset of high-level properties more generally, then it is reasonable to conclude that we can perceptually represent at least some high-level properties. This result strengthens the case for a richer conception of perceptual content and has important implications for ongoing debates in the philosophy of perception concerning the scope and structure of perceptual representation.

Elodie Boissard: *On Strength of Desires*

The notion of strength of desires is used in naive psychology, which explains behavior through beliefs and desires that generate intentions that cause actions. It is also used in the philosophy of mind, which studies beliefs, desires, emotions, and other types of mental states and processes. Finally, it plays an important role in moral philosophy, in central debates in each of its subfields. Normative ethics, which mainly thinks about the good life, often discusses a hierarchy and discipline of desires, bringing their strength into play, as is clearly shown in the discussion on akrasia or weakness of will. In ethics applied to psychiatry, the question is raised of whether some desires are irresistible or not, in particular compulsive desires involved in pathologies such as obsessive-compulsive disorders or addiction. This issue is related to the debate on moral responsibility, in meta-ethics, because the irresistible strength of some desires could be a psychological factor providing excuses for perpetrated wrongs.

However, the notion of strength of desires is obscure. At first glance, it appears to be a quantification of desires, but various intuitions come into play when determining what this quantification applies to, drawing on as many possible theories of desires. It may apply to phenomenological intensity, to a causal power with regard to actions, to an influence on mental states (beliefs, other desires, emotions…), or to an order among the preferences revealed by an individual's choices. These approaches refer respectively to the theory of desires as dispositions toward pleasant states of consciousness (Strawson, 1994), to the theory of desires as behavioral dispositions (Smith, 1987, 1994), to the attentional theory of desires (Scanlon 1998), and to the theory of desires as guiding reward-based learning processes (Schroeder, 2004). This philosophical debate sometimes suffers from insufficient empirical input from psychiatry, psychology and neuroscience, even though these fields offer a wealth of recent relevant data. For example, numerous studies focus on cravings, namely desires that are unusually strong and now included in the clinical criteria for addiction (Addolorato et al., 2005; Bergeria et al., 2021; Flanagan, 2020; Redish & Johnson, 2007; Robinson & Berridge, 1993; Vafaie & Kober, 2022). Such studies lend themselves well to an empirically informed philosophical discussion on the strength of addictive desires (Lavallee, 2020b, 2020a; Pickard, 2024).

What, then, is strength of desires? Our reflection will proceed in two stages. First, we will examine philosophical theories of desires and argue that the attentional theory is most consistent with empirical data. Second, we will show that this theory provides a conception of strength of desires as a power to influence attention, which is also consistent with various data.

In the first part, we will first refute the theory of desires as dispositions toward pleasant states of consciousness, drawing on Robinson and Berridge's dissociation of the circuits of wanting and liking in their incentive sensitization theory of addiction (Robinson & Berridge, 1993, 2025). Secondly, we will reject the theory of desires as behavioral dispositions by drawing on the distinction made in psychology and neuroscience between four systems of action: reflexes, Pavlovian conditioning, habitual behavior, and goal-directed action, with only the latter system involving desires (e.g. Rangel et al., 2008). This fourfold division shows that many dispositions to act are not desires, and we will further argue that some desires, mainly impossible or unrealistic ones, do not dispose one to act. Thirdly, we will dismiss the theory of desire as guiding reward-based learning processes by emphasizing that these processes can be part of Pavlovian conditioning and habitual behavior. These latter may not involve any representation of the value of the object that serves as a reward, nor of the object itself or of obtaining it as the result of some actions, as proved in outcome devaluation procedures, whereas this representation is supposed to correspond to desire in this theory. Finally, after proceeding in this way by elimination, we will argue positively in favor of the attentional theory of desires by maintaining that it is consistent with the inclusion of desires in the goal-directed action system. In this system, a desire corresponds to a motivational state that causes actions together with the representation of their ends and

means by beliefs or other cognitive states. With the function of causing actions that are guided by cognitive states, this motivational state is capable of engendering a number of mental states by directing attention to certain stimuli.

In the second part, we will build our definition of strength of desires based on the attentional theory of desires in three stages. First, we will distinguish the strength of a desire from its phenomenological intensity, from its causal power with regard to actions, and from the priority of resulting preferences over other of the individual's preferences. Taking strength as the active power of something, we will argue that strength of desires lies in their power to influence attention, drawing on the attentional theory of desires and responding to certain objections.

Second, we will discuss the main competing theory, namely the theory of causal power with regard to actions, that is the orthodox philosophical theory about strength of desires (Schroeder, 2020). In particular, we will argue that a desire can be strong and yet devoid of causal power with regard to actions because it is deprived of it by a mental state other than a desire, such as a belief, or by mechanisms of repression, or because the psychological or physical conditions required to carry out relevant actions are not met. We will rely on analyses by Nordenfelt in the philosophy of action applied to irrationality and mental disorders (Nordenfelt, 2007). Thirdly, we will argue that diverse phenomena to which psychiatry, psychology and neuroscience refer as strength of desires, such as irresistibility of cravings, are better accounted for by the notion of a power to influence attention than by the notion of a causal power with regard to actions.

Alon Chasid: *A (new) argument for i-desire*

In this talk, I present a new argument for i-desires. I defend the thesis that, to explain emotional responses to fiction, i-desire—i.e., an imaginative analog of desire—must be posited. My argument is based on the fact that we respond emotionally to fiction not only on first engagement, but also in re-engaging with a work of fiction. It is a well-known fact that, in reading or watching a work of fiction, we respond emotionally to the depicted events as if they were real. E.g., in reading (watching) a work that depicts an impending disaster, we feel fear, stress, suspense, etc., although we know that no real disaster is imminent. This phenomenon gives rise to the paradox of fiction (see Tullmann 2024 for an overview): our emotional responses to fiction are puzzling, since normally we do not respond emotionally to events we believe do not occur. Indeed, the paradox assumes that, setting aside knee-jerk or pre-cognitive emotions, emotions are explained in folk-psychological terms, or at least sensitive to beliefs and desires. The now commonly-accepted solution of this paradox holds that: (1) we do not believe the fictional content, but imagine it, and: (2) attitudinal imagining—the cognitive state that arises in engagements with fiction—can generate emotions in much the same way as belief with the same content does (see, e.g., Arcangeli 2018, ch. 2; Chasid 2021). That is, folk-psychology must be revised: not only belief, but also imagining, can function as the cognitive basis of emotions.

Accepting this solution, I focus on a related question: which desire must be paired with imagining to complete the folk-psychological explanation? Two kinds of desires are relevant here, both are problematic. First, a desire concerning how the work of fiction unfolds—e.g., that according to the work, no disaster will happen—cannot explain the emotional responses in question. Fear or stress (e.g.) cannot be explained by the desire that the work not depict an impending disaster. Desiring that the work not say what it does, or not unfold as it does, might generate aesthetic displeasure, disappointment with the author, etc., but not fear or stress; the latter emotions normally pertain to the depicted events, not to the fact that the work depicts them. Second, a 'direct' desire regarding the fictional event itself—e.g., a desire that no disaster will happen—is also problematic. Although certain philosophers accept this option (e.g., Kind 2013; Spaulding 2015), others reject it mainly on the grounds that we do not normally have desires about (what we believe to be) non-existent

objects or events (see, e.g., Currie and Ravenscroft 2002; Eagen and Doggett 2007; 2012). The latter argue that, since no desire can explain emotional responses to fiction, a desire-like imagination—an i-desire—must be posited to explain them. That is, we feel fear and stress since we imagine an impending disaster, and i-desire that it not happen. At bottom, the debate over i-desire is about the minimal requirements for a mental state to be deemed a desire. 'Conservatives' argue that the profile of desire can accommodate the abnormalities of emotions in fiction, whereas 'liberals' argue that it cannot, hence i-desires must be posited. My argument falls under the latter rubric.

The argument I propose focuses on the satisfaction conditions of desire. Setting aside minor complications, a desire that p is satisfied if and only if p. Of course, one can desire that p without knowing that p, thus without knowing that one's desire is satisfied; similarly, one can falsely believe that p, i.e., believe that one's desire is satisfied, when it is not. Importantly, the psychology of desire depends on our beliefs about its satisfaction conditions. Suppose one desires that p; if one comes to believe that p fully obtains, one's desire, as well as the emotions it generates, tend to vanish or become non-occurrent. E.g., I desire to avoid a certain surgery; this desire, together with the belief that I must undergo the surgery, renders me anxious, upset, etc. I then come to believe that I don't need any surgery. Consequently, my initial desire (which I now believe to be satisfied) vanishes, and so do my anxiety and dismay. Now consider re-engagements with fiction. Enjoying a work, we sometimes choose to re-engage with it. In doing so, we can respond emotionally in ways similar to how we responded on our first engagement. When reading the work's early chapters, specifically those depicting an impending disaster, we feel suspense, fear, stress, etc.; and when reading the final chapter, which reveals that no disaster was ever imminent, we are relieved and pleased (Smuts 2021; note that psychological experiments confirm this fact: see, e.g., Green et al. 2008, Chun, Park & Shi 2020). The problem is that no desire can explain these emotional responses. For in re-engaging with a work, we know that the putative desire—e.g., that no disaster will happen or is even impending—is fully satisfied. Conservatives are forced to hold that, despite our knowledge that our (putative) desire is fully satisfied, we nonetheless continue to have it: we eagerly desire that the imminent disaster not happen, and hence continue to feel fear, stress and suspense, all while knowing that the desire's satisfaction conditions obtain. Since no desire behaves this way, I conclude that, to explain the emotional responses in question, we must posit i-desires.

On my view, i-desires are mental states that, despite being capable of generating emotions (when paired with imaginings), are not responsive to beliefs about any satisfaction conditions. More precisely, i-desires do not have satisfaction conditions: no real-world fact can satisfy them. Like imaginings, i-desires are mere simulations (in a very specific sense of this term): just as imaginings do not have (real-world) correctness conditions, i-desires do not have (real-world) satisfaction conditions. Nonetheless, when suitably paired with imaginings, i-desires can generate emotions, just as belief and desire do in non-imaginative contexts. I conclude by discussing some implications of this account of i-desires.

Thursday 2nd July 17:00 — Consciousness & Phenomenology (Kerkzaal)

Johan Chung: *Revisiting Conscious Access: Why Internalism Is Not Mandatory*

What is conscious access? Who/what is accessing? What is being accessed? The standard answers would say that conscious access happens when information borne by representational vehicles processed by our sub-personal systems becomes conscious by being accessed by a conscious subject. But do we need to accept the standard answers? And, perhaps more pressing, how do we know that the standard answers are those that will lead us towards our best hopes for a scientific theory of consciousness? I aim to explore these questions, and in doing so provide

grounds for doubting that the standard answers to our basic questions about conscious access are ones we all ought to accept without question.

First, we need a principled way of distinguishing what kind of explanation we should demand of a scientific theory of consciousness to provide. Once we have an idea of what kind of explanation we are after, we can think of a general methodological framework for how we might go about providing such explanations. Such a methodological framework aims to secure "attachment points" between our ordinary, first-personal reflections on consciousness with sub-personal psychological explanation. With these in place, we can begin an analysis of the motivations for positing conscious access.

Such analysis reveals why the standard answers to our basic questions about conscious access would seem attractive. However, the analysis also reveals that our standard answers do not necessarily follow from the general methodological framework, nor from a principled understanding of the desideratum of a scientific theory of consciousness. In light of this, the standard answers to our basic questions about conscious access are best understood as theoretical 'ground assumptions'. However, once we recognize this point, it becomes apparent that such 'ground assumptions' are but a few of many possible ground assumptions that could have been made instead.

From this perspective, I explore whether there are principled reasons flowing from our desideratum for a scientific theory of consciousness and the general methodological framework which would necessarily preclude a somewhat 'radical' alternative to our standard answers. This radical alternative conceives of conscious access in terms of a subject's direct access to environmental features, rather than access to representational vehicles in the head. Though such a notion has been widely dismissed in sub-personal theorizing as a matter of contingent historical fact, our analysis reveals that neither the desideratum for a scientific theory of consciousness nor our general methodological framework provides grounds for such a dismissal.

Finally, I explore some of the implications that taking on alternative 'ground assumptions' to our standard answers to basic questions about conscious access may have on current theorizing. In conclusion, which set of ground assumptions we ought to accept cannot be decided by the general aims and method of a scientific theory of consciousness. Neither should we leave the decision up to contingent matters of historical fact. What ground assumptions we make should rather be at least in part constrained by the explanatory power of these assumptions in addressing the kinds of questions that we are demanding answers to. Whether the standard answers to our basic questions about conscious access are our only available answers is a matter that cannot be decided until the alternative options have been explored thoroughly.

Leonard Dung: *Integrating mammalian and non-mammalian consciousness research: Metamodels and instantiation models of consciousness*

My aim is to identify and resolve a tension in contemporary consciousness science. On the one hand, mainstream work on the neural basis of consciousness—driven largely by studies of humans and other mammals—often treats processes in the cortex as crucial, and perhaps even necessary, for conscious experience (Malach, 2022; Michel, 2022). On the other hand, comparative work on non-mammalian animals—especially fish and invertebrates—has amassed increasing behavioral evidence that many such animals are conscious despite lacking a neocortex (e.g. Crook, 2021; Gibbons et al., 2022). Taken at face value, these two streams of research generate an uncomfortable choice: either downgrade the neuroscientific case for cortico-centrism, or downgrade the behavioral case for widespread non-mammalian consciousness. The guiding thought of this paper is that we should resist that forced choice. The methodological aim is not to

dissolve the tension by rejecting one side, but to develop a framework in which both kinds of evidence can retain their probative force while genuinely constraining theorizing about consciousness. Proposals that assign different neural substrates to different “levels” of consciousness can appear to dissolve the tension (Newen & Montemayor, 2023): perhaps subcortical processes suffice for a minimal form of awareness while cortical broadcasting is required for richer, more cognitively integrated consciousness. The paper argues that this move often underdelivers, because it does not answer the real point of friction: which processes are sufficient for phenomenal consciousness, and how do we test that sufficiency? Merely noting that there are fast subcortical control loops and thorough cortical integrations does not by itself reconcile evidence for and against non-cortical consciousness. Without a principled bridge between levels and the evidential roles they are meant to play, a two-tier theory risks collapsing into either a cortico-centric view (if the “minimal” level is not genuinely phenomenal) or a non-cortical view (if it is). Third, two broad reconciliation strategies in the literature—call them generality and distinct realization—each capture something right, but each also faces a serious obstacle when taken alone. The generality strategy seeks a highly abstract functional characterization of the consciousness-enabling architecture: one property F that is realized by the neocortex in mammals but by different anatomical structures elsewhere. This promises unification across species, yet it invites the familiar worry that overly abstract functional conditions become too permissive (the “small network” problem) and too thin to support mechanistic explanation. Conversely, the distinct realization strategy emphasizes that consciousness may be implemented by very different mechanisms in different taxa, so mammalian neural constraints need not generalize. But this threatens to sever the mutual constraints that make comparative research scientifically fruitful. If mammalian and non-mammalian research proceed in near-independence, it becomes unclear how either can robustly inform the other.

The paper’s positive proposal is to combine what is right in both strategies by explicitly developing two linked model types: 1. Metamodels are taxon-spanning models stated in abstract functional terms. They aim to capture the shared organizational features of whatever realizes consciousness across the range of conscious animals (or across a large subset of them). Metamodels prioritize unification: they articulate what kinds of roles, interactions, or architectures must be present somewhere in a conscious system, even if different species implement those roles with different anatomical parts. 2. Instantiation models are taxon-specific mechanistic models. They aim to specify the particular neural components, activities, and organization that realize consciousness in a given group (e.g., mammals, birds, basal vertebrates, arthropods, cephalopods). Instantiation models prioritize explanatory depth: they should support interventionist “what-if-things-had-been-different” counterfactuals and connect consciousness claims to concrete neural mechanisms.

Crucially, the proposal is not merely to have both kinds of models, but to treat them as mutually constraining. Metamodels guide instantiation-model construction by indicating what functional structures to look for when mapping candidate mechanisms in a new taxon. Instantiation models, in turn, constrain and refine metamodels by revealing which features are genuinely shared across taxa and which were mammal-centric artifacts of an initial model. Integration is therefore iterative: theorists should expect a co-evolution of increasingly detailed instantiation models and increasingly taxon-general metamodels, with adjustments driven by familiar scientific virtues—explanatory power, predictive success, and unification.

To demonstrate that this is more than a terminological reshuffle, the paper develops a worked example centered on the relationship between Unlimited Associative Learning (UAL) (Ginsburg & Jablonka, 2019) and Global Neuronal Workspace Theory (GNWT) (Dehaene, 2014). UAL functions as a promising metamodel: it characterizes a minimal architecture that supports flexible, open-ended associative learning through interactions among sensory processing, valuation, memory, motor control, and a central associating workspace-like hub. GNWT functions as a

paradigmatic instantiation model for mammals: it explains conscious access in terms of global broadcasting enabled by long-range cortical connectivity and workspace dynamics distributed across cortical networks. On the proposed reading, the point is not to force identity between UAL and GNWT, but to show how they can productively relate. UAL provides a taxon-general template—an abstract organizational profile that consciousness-realizing systems should approximate—while GNWT specifies one mammalian way of implementing that template with cortico-thalamic circuitry. I will illustrate the virtues of this methodology in recourse to a case study regarding comparative neuroanatomical work on basal vertebrates by Zacks and Jablonka (2023). They can be understood as asking: Where, in early fish brains, do we find structures that play the integrating and broadcasting roles specified in the UAL metamodel, and what modifications to the metamodel are required by putative this non-mammalian instantiation of consciousness?

Overall, the framework respects mammalian evidence suggesting that particular cortical structures are deeply implicated in human consciousness, while also leaving space for non-mammalian consciousness supported by sophisticated behavior. It does so without making consciousness either trivially easy to realize (pure generality) or radically disunified (pure distinct realization). More broadly, it offers a concrete methodological blueprint for comparative consciousness science: build metamodels that unify across taxa; build instantiation models that explain within taxa; and let each constrain the other in an explicitly iterative research program.

Robyn Carston: *Metaphor: Non-propositional effects and the communication of 'what it's like'*

Accounts of the comprehension of metaphorical language in linguistic pragmatics tend to focus on propositional contents; this is especially apparent in the discussions of Grice (1967) and Searle (1979) who develop accounts where 'what is said' is patently false/uninformative, hence not communicated/meant, but is a vehicle for the communication of implicatures (implicitly communicated propositions) which constitute the speaker's meaning and which hearers infer in order to preserve the presumption that the speaker is observing the cooperative principle. There is no mention here of any non-propositional effects, i.e. imagery, attitude, affect, even though, intuitively at least, the examples they use ('You are the cream in my coffee', 'Sally is a block of ice', 'Juliet is the sun') do seem to have such effects. Even the more cognitive-scientific approach of Relevance Theory (Sperber & Wilson 1995, 2008) has tended to focus on propositions, suggesting that: 'What look like non-propositional effects associated with the expression of attitudes, feelings and states of mind can be approached in terms of weak implicature' (1995: 222).

At the other extreme, some philosophers of language have maintained that there is no such thing as metaphorical meaning, that the only propositional content metaphors have is their literal meaning, and that what a metaphor does is make us 'see' the topic in a new way, prompting open-ended responses in us of an imagistic sort (Davidson 1978). Metaphorically used language triggers an imaginative engagement which lies outside the purview of a systematic pragmatic theory (Lepore & Stone 2015); the kind of imaginative activity at work here has an essentially private significance, and even if interlocutors sometimes end up sharing insights by such means, these insights are not components of the shared meaning-making endeavour. The faculty at work here is taken to be quite distinct from our standard semantic/pragmatic (proposition-generating) capacities.

I have tried to develop an account which is both more inclusive and more balanced than either of these two extremes (Carston 2010, 2018, forthcoming), making space for both the propositional and the non-propositional, which are differently weighted in different sorts of cases (more weight on the propositional in relatively conventional cases and those used for pedagogical purposes, and more weight on the imagistic/affective in creative, extended or poetic cases). Metaphorical

language use is, after all, a very heterogeneous phenomenon, varying along several dimensions (familiarity/novelty; lexical/extended; spontaneous/crafted). Neuroscientific evidence shows that degree of activation of various sensory-motor brain areas depends on degree of familiarity of a metaphor; the less familiar the metaphor, the higher the activation (Desai et al. 2011; Ospina et al. 2024). Although multimodal simulations are usually thought of as unconscious (sub-personal) processes, in cases of novel metaphor requiring more time and effort for comprehension, these simulations may become consciously available and entertained (scrutinised, manipulated and enjoyed) as what literary theorists think of as 'mental imagery'.

The questions now are why we use language metaphorically, what is gained by prompting conscious simulation/imagery in our interlocutors, whether this is just a byproduct of the kind of mental processing required for propositional understanding, or it can have some sort of epistemic value in itself. Scientific/pedagogical metaphors (e.g. Infectious diseases described as warfare between an invading army and the defences of the organism) do seem to have an epistemic role: "We exploit our knowledge about warfare (the source domain) in order to conceptualize and thereby to come to better understand the notion of an infectious disease (the target domain)." (Kompa 2021: 38). What about more personal non-scientific metaphors, where we attempt to communicate subjective experiences (sensory-perceptual, emotional/affective, attitudinal)? In a recent paper, Kind (2024) discusses the difficulties we have describing our own phenomenal states; as she puts it 'we lack an adequate subjective vocabulary for describing phenomenology' (p.119). She suggests that one of the ways by which we may make some progress in overcoming this 'impoverishment problem' is through the metaphorical use of language. Examining a corpus of utterances by patients attempting to describe various states of pain or illness that they are experiencing, she shows how they almost inevitably turn to metaphorical language (e.g. rheumatic pain is described as 'a sort of bubbling').

I develop this idea here, arguing that one of the central uses of metaphor is precisely to overcome the limitations inherent to language in the literal expression of phenomenal states, something that gets its most heightened and evocative use in more crafted (often literary/poetic) metaphors. Consider, for instance, the following lines (from the poem 'Hana' by Oskar Davico), where the poet gives expression via metaphor to the powerful phenomenal state, the rush of extreme (even contradictory) feelings, of being in love (all the more evocative in the context of the complete poem, which is metaphorical from beginning to end): Love is so lonely and so full of people. Love is the lighthouse and the rescued mariners.

Rather than accessing files of propositional information stored under specific concepts (here, LONELY, FULL OF PEOPLE, LIGHTHOUSE, RESCUED MARINERS), it seems more plausible that, as Pilkington (2022), a strong advocate of a non-propositional account of poetic metaphor, puts it: '… we are encouraged down the route of accessing phenomenal memories, which, according to Damasio (1989), are also attached to concepts, to construct a phenomenal state that is affective as well as perceptual.' The hearer/reader with the appropriate understanding of the phenomenon literally described by the metaphorical vehicle language can draw on that experience (whether sensory-perceptual and/or emotional-affective) and apply it to the metaphor topic (here 'love') so as to reach an understanding of the experience that the speaker/writer is trying to express/communicate (Camp's (2006) view of metaphor as demonstrative-like is similar).

In this respect, then, such (non-scientific) metaphors can be said to have a certain kind of epistemic value: they enhance one's grasp of someone else's phenomenal state, perhaps enabling better understanding of one's own comparable phenomenal state and so achieving a sense of the sharedness of an experience which had previously felt to be wholly subjective, ineffable, and 'locked within'.

Joulia Smortchkova: *The feeling of having a first impression*

Cognitive feelings are often thought to play a role in cognition when we encounter entities and appraise them as familiar or unfamiliar. With respect to these issues, the feeling of familiarity (and more rarely, the feeling of unfamiliarity) has so far been at the center of discussion. Philosophers and psychologists describe the feeling of familiarity as the subjective sense that we have encountered something before, even if we cannot immediately retrieve the information relevant to recognizing that entity. Psychological models construe the feeling of familiarity as a memory signal that arises during a search for a match in memory for a perceived entity, with expectations playing a prominent role (Whittlesea & Williams, 2000). By contrast, when we immediately recognize an entity, we experience recognition, in which case there is little or no feeling of familiarity. The feeling of familiarity is characterized by several features: it is category-dependent (an object may be familiar with respect to one category and unfamiliar with respect to another), it is partially epistemically opaque (we lack access to the entire web of relevant information, even if some information about the entity is available to us), and it is gradable, in the sense that it can be modulated by expectations (Dokic, 2025, Chapter 7). Cognitive feelings are usually considered metacognitive because they monitor and regulate first-order cognitive processes, such as a memory search in the case of the feeling of familiarity. Metacognitive feelings are used as cues to guide one's mental actions, for instance by prompting one to continue searching for missing information.

This talk focuses on what happens when we encounter entities that are not familiar to us—cases in which we have neither recognition nor a feeling of familiarity. While the most obvious response might be that we experience either an absence of familiarity or perhaps a feeling of unfamiliarity (Dokic, 2025, Chapter 7), I will argue that the feelings involved in these experiences are more complex and more informative than they prima facie appear. More precisely, I will show that there is a richer kind of experience, which is part of our everyday phenomenology but has not yet been explored by philosophers of feelings: the feeling of having a first impression. To illustrate with a few examples: we step outside London's St Pancras station for the first time and immediately like the city, or we meet a new colleague and instantly judge her to be nice. These immediate assessments are reflected in everyday talk, where the notion of a "first impression" is used to make sense of our experiences. There has been some theoretical interest in what happens when we encounter new entities. In such cases, apart from experiencing a feeling of unfamiliarity, as mentioned above, we might experience a feeling of novelty (Weierich et al., 2010), a feeling of curiosity (Goupil & Proust, 2023), or perhaps a feeling of surprise (Reisenzein et al., 2019). However, I contend that these experiences are less frequent than the experience of having a first impression of a newly encountered entity. My main goal is to offer a positive characterization of the feeling of having a first impression. I will argue that this feeling is a complex state composed of metacognitive awareness, an affective reaction, and a distinctive phenomenological profile. Like the feelings of familiarity and unfamiliarity, the feeling of having a first impression can be category-dependent (we can have a first impression of an entity we already know if we encounter it under a new category), epistemically opaque (in the sense that we do not know which information the feeling is based on), and gradable (in the sense that it is modulated by expectations, which may affect its intensity) (Kind, 2021).

However, this feeling is more complex and does not merely track information such as "this entity is new," "unfamiliar," or "unexpected." I will show that the feeling of having a first impression is best understood in terms of its function in cognitive life. It appears to play a central role in enabling a rapid, valenced, here-and-now assessment of a newly encountered entity, even when little information is available. In this respect, although the feeling of having a first impression may track the subject's first-order cognitive processes (as other cognitive feelings do), it is interpreted (Unkelbach, 2006) as being about the entity in front of the subject. That is, the affective information

is used to evaluate the new entity with respect to its importance and relevance to one's current goals (Zadra & Clore, 2011). This suggests that, in some cases, cognitive feelings are outward-directed and function to help us navigate the world around us, rather than only to monitor and control our mental processes (Proust, 2013). Introducing the feeling of having a first impression into the domain of cognitive feelings will hopefully be significant to understanding how the mind deals with novelty.

Thursday 2nd July 17:00 — Memory & Agency (Grote zolder)

Ryan Mokhtari and Judith Carlisle: *Behavioral modernity through the integration of autonoesis and metacognition*

Animal studies suggest that neither autonoesis (mental time travel) nor metacognition (thinking about thinking) is unique to humans. Their integration, however, may have uniquely contributed to the cognition of behaviorally modern humans. This autonoetic-metacognitive merger was likely mediated by increasing social complexity and operated through two mechanisms. First, applying metacognition to reciprocal interactions may have led to the formation of the concept of self as an objective reality. Second, complex societies likely expanded semantic knowledge to include abstract concepts, most importantly the concept of time. Unlike the phenomenal experiences of self and time, understanding their meanings and implications must have transformed human cognition, most clearly manifested in ritual burial during behavioral modernity. The expansion of semantic knowledge, and its necessity for complex cognition, is supported by comparative neuroanatomy showing that semantic brain areas evolved more recently than those associated with autonoesis and metacognition.

Thor Grünbaum: *Memory, Reinforcement Learning, and Temporally Extended Agency*

This paper is about the role of memory in temporally extended intentional agency. An agent engages in temporally extended intentional agency if they intentionally pursue a single goal persistently across time (for instance, taking half a day to assemble a complicated piece of IKEA furniture). Depending on how one conceives of the way in which agency extend through time, one will conceive differently of the roles played by memory in enabling agency to extend.

We find two different conceptions of temporally extended intentional agency in contemporary neuroscience, psychology, and philosophy. On the one hand, some people think that temporarily extended agency is reducible to a series of proximal decisions (Sripada, 2025). In this paper, I focus on recent computational accounts of action selection and control in Reinforcement Learning (RL) models in cognitive neuroscience (Niv, 2009; Collins, 2024). According to this proximal conception, an agent's practical decision is always a decision about what to do at the most immediate choice point. The shape of a long-term project is constituted by the way in which the stable shape of the reward landscape shapes the trajectory of in-the-moment choices.

On the other hand, some people think that temporally extended agency can be controlled by long-term intentions to do something in the distal future (Bratman, 1987). According to this distal conception, an agent can make future-directed decisions and thereby form intentions that control actions by their temporally distal goals. The shape of a long-term project can be caused by a stable, persisting intention formed by making a future-directed decision. The stability of the intention can at times be at odds with proximal values and rewards.

The proximal and distal conceptions of temporally extended intentional agency imply different conceptions of memory. According to the proximal conception, decisions about what to do are always about imminent choices, and memory plays an important role only by providing the decision-maker with information about the world. Semantic and episodic memory helps the agent

build a cognitive map of the choice space (Mattar & Daw, 2018). The map represents values and facts. Memory thus helps the agent calculate possible cumulative future rewards given choices made here and now (Shohamy & Daw, 2015; Gershman & Daw, 2017). The stability of temporally extended agency is partly explained by a stable cognitive map (i.e., a stable space of options with stable assignments of values and possibilities) and memory mechanisms to maintain and retrieve the map.

By contrast, according to the distal conception, decisions about what to do can be about actions in the future, and memory can play a role in retaining a representation of the intended future action (Grünbaum & Kyllingsbæk, 2020; Oren et al., 2021). Future-directed intentions are sticky. Once formed, the intention is not reevaluated at each new choice point. Memory plays an important role in shielding the agent from in-the-moment motivational fluctuations and temptations.

In this paper, I argue that semantic and episodic forms of memory are not sufficient for satisfying this additional role of memory in retaining a representation of and commitment to a future action. I argue that, given their reliance on only semantic and episodic forms of memory, recent RL versions of the proximal account of agency are unable to explain everyday cases of resisting temptation, familiar from the planning theory (Bratman, 1987; Holton, 2009).

My argument proceeds as follows. First, I describe a simple case of an agent resisting the temptation of postponing the performance of a costly action. I consider an agent who must perform a costly action before Friday but on each weekday Monday to Thursday has the option of delaying the action. As time approaches Friday, the risk of not performing the action (due to unforeseen events) increases. Psychological studies have shown that, in such situations, agents would be able to plan and execute the action early in the week (Sheeran et al., 2024).

Second, I introduce the basic logic of RL models of temporally extended intentional agency. My focus is on intentional action. Consequently, I will only discuss so-called model-based accounts of action selection (Daw & Dayan, 2014). I provide some details on how current RL models conceive of episodic memory as providing information about the possible rewards to be obtained by different possible actions.

Third, I review three different ways in which an RL account might explain the ability to plan and execute the costly action early in the week. According to the first RL explanation, an agent's true preferences are revealed by the actions they select (Redish, 2013). Consequently, performing the action early in the week just demonstrates that it was in fact not costly (compared to other available options). According to the second RL explanation, if agent's compute the increasing risk involved in postponing the performance, in the long run the cost of postponing might be too high. The agent might adopt some kind of meta-control policy that will allow it to perform the action early in the week (Kool et al., 2018). According to the third RL explanation, agents can simulate performing the early action which could enable them to build so-called "successor representations" (Momennejad, 2020). This might enable the agent to chunk state-action trajectories. The main argument of this paper is that all reviewed RL explanations are unable to provide a satisfactory explanation of our ability to plan and later intentionally perform the costly action at a time when it can easily be postponed. The crux of the argument is centred on RL models' proximal conception of temporally extended agency and their commitment to semantic and episodic roles of memory.

I conclude the paper by briefly sketching an alternative conception of the role of memory in temporally extended intentional agency. If RL models with their proximal conception of action selection and control are unable to explain everyday temptation cases, we have reason to opt for

a more distal account of temporally extended agency. I briefly suggest a possible alternative role for memory on such an account.

Frederik T. Junker: *Beyond Deliberation: How Memory Shapes Intentions*

Consider an agent who forms an intention to pursue a particular career, judging the path well suited to her. Months pass. She never sits down to reconsider, and no decisive new information arrives. Yet when the plan next comes up it no longer moves her as it once did: remembered disappointments are more salient, the imagined future less appealing. What changed for her, and how does it bear on accounts that treat deliberation as the main driver of intention revision?

Leading theories of temporally extended agency offer only partial answers. Planning theory, in Bratman's tradition, treats intentions as conduct-controlling and reasoning-anchoring states, stable across time and revised through deliberation (Bratman, 1987; Holton, 2009). Valuationism presses a rival picture: action runs on value representations consulted at each decision, and intentions carry little explanatory weight (Sripada, 2025). The two seem to emphasize different timescales of action coordination. But neither fully explains the processes gradually affecting a standing commitment across the stretches when the agent is not deliberating.

I argue that this interval is not inert. Because intention must be stored in memory until the time to act arises, intentions are likely reworked offline by processes known to affect memory. Value-updating, through experience replay, modifies the expected value tied to the intention, and so its motivational pull (Liu et al., 2021). Affect-tagging modifies affective valence, altering both pull and accessibility. Semantization makes the content of memories more abstract and less episodically detailed over time (Aronowitz, 2025). None of these processes amounts to deliberation: each runs while the agent is at rest, accumulating in small increments and reaching no conclusion she reflectively avows. Two pathways should be distinguished: a change to stored representations count as a change in the intention when it alters the representations constituting the intention itself that are responsible for conduct-control or reasoning-anchoring; otherwise, it is a change around the intention – say, to the reasons on which the intention is based.

What determines when offline change reaches deliberation? I argue for a metacognitive controller that weighs the benefit of deliberating against its cost, engaging it only once certain signals – an affective marker, conflict, or shifts in value or content – cross a threshold. The agent registers the shift only when offline changes have been substantial enough to warrant engaging deliberation. This fits naturally with the machinery native to valuationism.

However, valuationism should also acknowledge the role of intentions in coordinating our activities over time in the pursuit of long-term goals and in reducing cognitive load. A standing intention filters: inconsistent options tend to drop out before their value is computed. A gate keyed purely to value cannot do this, since the excluded option may be one the agent would value highly. The filter also answers to consistency with our prior commitments.

Representational transformation re-keys this filter: as the commitment abstracts from a particular job toward a career of a certain shape, a comparable job elsewhere, once ruled out, becomes admissible – a change in consistency, not just value.

The result is a hybrid account: planning theory's farsighted intentions and valuationism's nearsighted value representations and updates become parts of one system, with incremental transformations within memory affecting the representations guiding our actions between deliberative episodes.

Alexandria Boyle and Eva Read: *Animals, mental time travel and the harm of death*

Would it harm you if you were to die in the prime of life? And would the answer be different, if you were a non-human animal? Several philosophers have answered yes to both questions, arguing that while a premature death is among the most serious harms that can befall the average adult human, nonhuman animals are not harmed by death, or are harmed by it significantly less than we are. This is because humans and animals are alleged to differ with respect to their capacities for mental time travel. Mental time travel is the capacity to mentally project oneself backward and forward in time, to recall experiences from one's past (a type of memory called 'episodic memory') and to imagine experiences one might have in the future. Mental time travel looms large in the experience of the average adult human. However, according to these arguments, animals either lack mental time travel altogether, or their mental time travel capacities are less rich than ours. So, animals are not harmed, or are less harmed, by death.

For example, Belshaw (2015) proposes that death can only harm individuals who have desires about the future. Animals are unaware of their future existence, and hence lack such desires and are not harmed by death. Similarly, Singer (2011) argues that preference utilitarians should view death as harmful to the extent that it frustrates an individual's future-oriented preferences. Singer proposes that many animals exhibit mental time travel, and hence may have future-oriented preferences. However, Singer suggests that animals probably have fewer, weaker preferences relating to the future, and are therefore harmed less by death than humans. McMahan (2015) argues that death harms an individual to the extent that they have a 'time-relative interest' in continuing to live. Having such an interest requires being psychologically connected to one's future self. He proposes that the inferior mental time travel capacities of animals leave them substantially less psychologically connected than humans – and hence substantially less harmed by death. And Nussbaum (2023, 2013) argues that death is harmful when it interrupts one's temporally extended projects, rendering one's investments in those projects vain and futile – and that having such projects involves mental time travel. Like Singer, Nussbaum is liberal-minded about the distribution of mental time travel: she proposes that it's to be found in all mammals and birds, but is unlikely to be found in fish. So, death is harmful to the former, but not the latter.

In this talk, I offer two critiques of these arguments. First, these arguments are hostage to empirical fortune, and their fortunes look unpromising. They rely on extremely controversial claims about animal cognition which do not survive contact with the empirical literature. The evidence about mental time travel in animals is subject to significant debate and uncertainty, to the extent that mental time travel cannot confidently be ruled in or out for any nonhuman animal – so, confident pronouncements the mental time travel capacities of any animal are misplaced. Whilst Singer and Nussbaum acknowledge some uncertainty and recommend precautionary reasoning – attributing mental time travel to animals whose possession of it is uncertain – I show that the application of precautionary reasoning to this question is vexed. This is because, if mental time travel is found in animals, we should expect it to exhibit interspecific variation (Boyle, 2022, 2024; Boyle & Brown, 2025; Schwartz & Boyle, 2025). That is, mental time travel is unlikely to behave in exactly the same ways, or perform all of the same roles, in nonhuman animals – and the more widely distributed it is in the animal kingdom, the more we should expect it to vary. The more interspecific variation it exhibits, the less confident we should be that it occupies the normatively important roles identified in these arguments – that is, underpinning desires about the future, the pursuit of long-term projects or psychological connectedness. So, whilst precautionary reasoning might point us toward a very liberal account of the distribution of mental time travel in nature, that liberal distribution would undercut the very motivations for adopting a precautionary stance in the first place.

Second, these arguments anthropofabulate (Buckner, 2013). That is, they tie the criteria for possessing certain cognitive traits, namely future-oriented desires, temporally extended projects

and psychological connectedness, to an exaggerated account of what it is for humans to have those traits – in particular, exaggerating the role of mental time travel. As such, these traits might be realised in animals who lack mental time travel capacities. When we adopt a more expansive account of what future-oriented desires, extended projects and psychological connectedness might look like in nonhuman animals, we find sparse but suggestive evidence. On this basis, I argue that we should take seriously the possibility that many animals are harmed by a premature death.

Thursday 2nd July 17:00 — Morality & Virtue (Spiegelzaal)

Leda Berio and Daniel Kelly: *Moral Psychology for World Travelers: Playfulness as the Virtue of Ethical Codeswitching*

Berio and Kelly (2025) calls for the creation of an ethics of code-switching that provides moral guidance for exercising the feature of selves they call multiplicity, the psychological capacity to occupy many social roles and bear a range of social identities. This paper takes up that call, using Lugones (1987) to argue that the virtue governing ethical code-switching is playfulness.

For Lugones, world-traveling is the activity of moving through social groups, crossing borders between different cultural "worlds" and adapting to the norms and values that differentiate them. Good world-travelers are what we would call code-switchers: "The shift from being one person to being a different person is what I call "travel." This shift may not be willful or even conscious … Even though the shift can be done willfully, it is not a matter of acting" (p.17). Her notion of playfulness expresses a positive vision for how world-travelers can do this well. It involves being curious, open to novelty and surprise, and flexible, willing and able to try on new roles and engage in self-construction.

We argue playfulness is the virtue of code-switching by articulating a general picture of virtue rooted in contemporary empirical moral psychology (Westra 2018, Doris 2022). Virtues are ideal character traits. They guide ethical actions and attune judgments, sensitizing them to the challenges and values of different domains. Virtuous people are disposed to conduct themselves in the happy medium between vicious extremes of deficiency and excess. Virtues are best represented by moral exemplars, individuals whose actions and judgments manifest the virtue in action, making them positive role models crucial to moral education.

Next, we fill in this template for playfulness, arguing its domain of activity is self-presentation in Goffman's sense (1959, Berstler ms): the way a person puts forward different social identities in different situations, tailoring their posture, status, dialect, conversational, and emotion expressions so they are appropriate to the social role they are acting from.

Vices of deficiency in self-presentation include traits like rigidity and self-seriousness. These manifest in failures to adjust to and harmonize with one's social environment, the inability or unwillingness to see the need to, the hardened belief that others should adapt themselves to you. They are exemplified in the awkwardness of the stiff, truculent inflexibility of the insistent blowhard, the strident assertiveness of the privileged jerk (Schwitzgebel 2020). (These are Lugones' "arrogant perceivers").

Vices of excess in self-presentation manifest in behaviors and judgments where one adjusts themselves to a situation too much. This can take several forms. Phonies, frauds, and imposters exhibit the trait of Machiavellianism. They are exceptionally in control of their self-presentation, but they use their honed code-switching skills to intentionally misrepresent themselves, concealing their true intentions and lying about their actual ends. Others lack control of their self-presentation, exemplifying the trait of vicious docility (Hare 2016). Such individuals are sensitive to the worlds they travel through, but are unable to regulate their influence. They are easily

overwhelmed by the power of situational factors, only weakly resistance to social pressure. We are all susceptible to these forces, but docile code-switchers lack the kind of self-possession (Sandel 2022) required to be genuinely playful in the face of them. They fall short of virtue by immoderately using their facility in self-presentation: spineless relationship chameleons who reinvent their identities with each new partner, overly eager people pleasers who reshape themselves for approval and status, servile submissives who obediently bend the knee to any assertive authority.

Our paper culminates with a positive account of playfulness weaves together several contemporary ideas. Dover (2022) shows how erotic curiosity allows others' interpretations of us to shape our own self-understanding and self-presentation. McGeer (2019) argues that good moral agency requires the ability to nimbly switch between egocentric and allocentric perspectives on oneself, assessing and adjusting to one's social surroundings while simultaneously maintaining a sense of one's place in them. Bessler and Oishi (2020) describe the well-being benefits that accompany psychologically rich lives that are abundant with shifts in perspectives, motives, and novel experiences. Nguyen (2026) shows how games capitalize on and nourish our psychological multiplicity, allowing us to joyfully slip in and out of different agencies. We add that virtuously playful code-switchers are well-equipped to avoid the pitfalls of gamification and value-capture, because they are well-equipped to resist what we call "identity capture", the phenomena of getting trapped by social influence, stuck in a single identity by the expectations of others, mistaking one facet of their self or one mode of their self-presentation for the whole thing.

:

Pascale Willemsen, Lucien Baumgartner and Nikolai Shurakov: *New Empirical Studies of Moral Praise*

Blame has long been under scrutiny in both moral philosophy and moral psychology, while praise was either neglected or assumed to be a merely positive moral analogue of blame (see Telech, 2022). Recent literature challenges this assumption on both theoretical (e.g., Anderson et al., 2020; Johnson King, 2025) and empirical grounds (e.g., Anderson et al., 2024; Bostyn & Knobe, 2025). Research on the blame–praise asymmetry increasingly promotes the view that praise plays a distinctive social role: it contributes to relationship-building, signals shared values, and communicates social norms (Stout, 2020; Anderson et al., 2020; Telech, 2022). Related empirical work in pedagogy and developmental psychology further shows that praise and reward are often more effective than blame and punishment in promoting prosocial behaviour, indicating that some functions traditionally attributed to blame may in fact be better served by praise (Johnson King, 2025, pp. 2–3). At the same time, much of the psychological literature focuses on positive reinforcement and rewards rather than on praise as a specifically moral evaluative practice. As a result, we still have only limited empirical evidence on how ordinary people understand praise. This paper addresses this gap by reporting the results of a study that addressed the question of who is in a position to praise or blame another, and on what basis.

A stranger at a playground who scolds someone else's child for misbehaving may be overstepping because they lack the relevant standing to blame. The literature on moral standing has identified the Relationship Condition: an agent lacks standing to morally evaluate another person if they lack a suitable relationship to that person. Applied to healthcare, this principle holds that, "unlike friends and relatives, doctors lack moral standing to praise or blame patients in health matters" (Varga et al., 2025, p. 665). Recently, Varga et al. (2025) reported surprising results: lay participants attributed greater standing to doctors than to friends. These findings challenge the Relationship Principle while leaving several questions unresolved. First, their study may have conflated "standing to give advice" with "standing to morally evaluate," even though these are conceptually distinct. A patient's health-related behaviour is clearly a doctor's business in an

advisory sense, but whether it is also the doctor's business to praise or blame the patient is a further question. Second, it remains unclear whether the results generalize across health conditions or are limited to obesity, the only risk factor examined in the study. Widespread stigma surrounding obesity may have influenced participants' judgments.

To address these gaps, we conducted a study employing a 2 (Advisor: doctor vs. friend) × 2 (Adherence: follows vs. does not follow) × 3 (Risk Factor: hypercholesterolemia, hypertension, sarcopenia) between-subjects vignette design. The study separates standing to give advice (DV1–DV3) from standing to morally evaluate (DV5–DV7), the latter decomposable into standing to praise within the Follows condition and standing to blame within the Does Not Follow condition. We also measure participants' own moral evaluations (DV4: whether the protagonist is more praiseworthy or blameworthy).

Our study (n = 504) shows that doctors are judged to have greater standing to give advice (Doctor M = 6.67 vs. Friend M = 3.52). Both doctors and friends are seen as having standing to praise (M = 6.06 vs. M = 4.83), although doctors' standing is higher. A different pattern emerges for blame: blame from a doctor is considered acceptable (M = 4.41), whereas blame from a friend is not (M = 2.75). Participants judged John (the protagonist) highly praiseworthy when following a doctor's advice and only somewhat praiseworthy when following a friend's advice (M = 3.82 vs. M = 1.63). Moreover, John is judged more blameworthy for ignoring a doctor's advice, but not blameworthy for ignoring a friend's advice (M = −3.28 vs. M = −0.84). The mediation analysis reveals that standing to evaluate is fully mediated by standing to give advice, providing an alternative explanation of the findings reported by Varga et al. In sum, our results provide novel empirical evidence concerning who is in a position to praise and how praise and blame work.

Markus Kneer and Juri Viehoff: *Fair Chances vs. Better Outcomes in Complex Moral Trade-Offs*

Allocation decisions in healthcare and public policy often pit “fair chances” against “better outcomes”: When two parties have equally strong claims to avoid a serious harm, an allocator can either decide determinately in the way that secures a slightly better overall outcome, or use a lottery to give each claimant an equal ex ante chances. Standard cost–benefit/cost-effectiveness analysis and many consequentialist theories treat fair chances as having no independent value and therefore recommend always choosing the outcome-maximizing option, even when the improvement is small. By contrast, non-consequentialist proposals—especially “partially aggregative” views (e.g. Voorhoeve 2014, 2016; Tomlin 2017; Mann 2022) and Kamm’s notion of “irrelevant utilities” (1993, 2008)—theorize that some small outcome improvements may be treated as morally disabled in the presence of weighty competing claims, so that fair-chance procedures are morally required. Recent work has confirmed the widespread prevalence of such complex non-aggregative intuitions amongst laypeople (Luptakova & Voorhoeve 2023; Kneer & Viehoff 2023, 2025). However, what is missing for a more complete theory of actual moral decision-making in this field is systematic evidence about (a) how sensitive such judgments are to the magnitude of the foregone gain, (b) whether they persist once choices are embedded in institutional roles, (c) how they interact with baseline need/priority differences, (d) whether they extend to non-harm goods, and (e) whether they survive explicit reflection about the strength of claims. Extending existing findings along these dimensions is both practically relevant for real-world decision-making beyond a narrow set of cases in bioethics, and of theoretical importance for an improved understanding of the various factors that shape moral judgment in this domain.

We report a set of vignette experiments ordered by increasing complexity: a simple harm-based rescue case (“Snake”), an institutional healthcare allocation case (“Ambulance”) that varies

whether additional benefit accrues to a separate person or to the primary beneficiary, and a non-harm/non-welfare case (“Drone”). In the presentation, we might also include a series of further experiments (total N=2745), which replicate our findings across contexts, formulations and experimental designs. Study 1 (“Snake”; N=309) adapts a classic two-claimant rescue case. Participants chose between saving one endangered person, saving the other endangered person while also producing an additional benefit for a third party, or flipping a coin between the two primary claimants. We manipulated the magnitude of the third-party benefit (“mild headache” vs. “tolerabilis”) and perspective (first-person vs. third-person). Added utility had a strong effect on choice ($\chi^2(2)=34.93$, $p<.0001$), while perspective did not ($\chi^2(2)=3.15$, $p=.207$). When the added benefit was small (“mild headache”), the coin flip dominated (65.4%), consistent with treating the extra benefit as an “irrelevant utility”; when the added benefit was larger (“tolerabilis”), the outcome-improving determinate option became the majority choice (58.8%), while coin flips dropped to 36.6%.

Studies 2–3 (“Ambulance / Emergency Operator”; N=323) embed the same structural choice in a healthcare allocation role and test interactions with priority/need. Participants read about an emergency-response shift leader who must direct a single ambulance to one of two simultaneous accidents. In both accidents, the driver will become paraplegic unless treated in time. Accident B additionally involves either a marginal injury (a bruised ankle) or a substantial injury (loss of a finger). Crucially, the extra injury either affects a second person (multiple-beneficiary institutional allocation) or affects the driver themselves (intra-personal need variation). In a multinomial model, both contrast and distribution significantly affected choices (contrast $\chi(2)=26.31$, $p<.0001$; distribution $\chi(2)=33.41$, $p<.0001$), while the interaction was not significant ($\chi(2)=1.20$, $p=.5494$). In the multiple-beneficiary/marginal contrast condition, about one in two participants chose the coin flip (exceeding both the other options and chance set at 1/3). When the marginal benefit increased to a substantial benefit (loss of a finger), the majority (about 60%) chose the outcome-maximizing determinate option, though a sizeable minority (about one in three) still chose the coin flip. When the substantial additional harm threatened the same individual (single-person/substantial), nearly everyone prioritized sending the ambulance to that individual rather than flipping a coin—showing that sufficiently large need differences can override fair-chance reasoning.

Study 4 (“Drone”; N=207) tests whether “irrelevant utility” judgments extend beyond welfare and health to cases where the “better outcome” is the preservation of impersonal/non-health value. Participants chose where to shoot down an imminent terror drone, knowing that five civilians would die whichever option is chosen; the only difference was whether the civilians were in an ordinary building or a “special” one (minor contrast: beautiful garden; moderate contrast: landmark building). Contrast significantly altered choice patterns ($\chi^2(2)=30.26$, $p<.001$). Aggregating responses into randomized (coin flip) vs determinate (either building), determinate choices were a minority in the minor contrast condition (46% against coin flip) but dominant in the moderate contrast condition (82% against coin flip). Importantly, the explicit claim-strength question showed that in both contrasts the vast majority (>91%) judged the two groups’ claims to be equally strong, even when they nonetheless favored the determinate, outcome-preserving option—suggesting that many subjects do not rationalize determinate choice by downgrading one group’s claim.

Across contexts, the data support a complex picture of non-consequentialist moral judgment: fair ex ante chances have independent moral force; many people treat some gains—small welfare gains in rescue/healthcare cases and modest impersonal goods—as too small to defeat the moral force of equal ex ante treatment; once the foregone gain becomes larger, many participants switch to the determinately better outcome. Moreover, priority to need can dominate when need differences are substantial and intra-personal. These patterns fit neither an unrestricted maximization rule nor a categorical “always lottery when claims are equal” judgment. They instead

point toward thresholded or partially aggregative accounts on which (i) chances matter in their own right, (ii) some benefits can be "disabled" in the presence of weighty competing claims, but (iii) this disabling is not absolute and can be overcome by weighty countervailing considerations.

Veronika Luptakova, Alex Voorhoeve and Matteo Galizzi: *Fallible but improvable: Repeated evaluation reduces moral inconsistency*

Extensive empirical research has documented inconsistencies in people's initial moral case judgments, which often result from order or framing effects and various cognitive biases. How people address such inconsistencies when given an opportunity to reconsider their case judgments has been less explored. Here, we examine whether people reduce inconsistency in their judgments on morally equivalent pairs of cases when given an opportunity to respond to cases again and in which of three settings they are most likely to do so: (i) repeated separate evaluation; (ii) joint evaluation; and (iii) joint evaluation preceded by an explicit prompt to assess their own consistency.

Using two exploratory studies – Pilot (N = 148) and Study 1 (N = 451) and a confirmatory online experiment Study 2 (N = 1,313), we find that participants tend to change their initial moral judgments to make them more consistent and that they are more likely to reduce their inconsistency under joint evaluation, when cases are presented together on one screen than under repeated separate evaluation, when they are presented in sequence on separate screens. However, we did not find strong evidence that the addition of an explicit consistency prompt – showing participants their own initial judgments and asking them to assess if those judgments were mutually consistent – further reduces inconsistency. While we observed that people who openly recognised their own inconsistency had a stronger tendency to reduce it, open recognition of inconsistency was neither sufficient nor necessary for reduction in inconsistency.

Our findings show that while people's initial judgments about morally equivalent cases are often inconsistent, people have a propensity to reduce their initial inconsistency when given an opportunity to reconsider their responses, even within the space of a short online experiment, and across different settings. This aligns with accounts of moral psychology, such as the moral consistency reasoning model, that posit that people are motivated to "treat like cases alike". This opportunity is most effective when inconsistency is made easy to spot by presenting cases together. Our results also suggest that people's moral judgments should not be dismissed as unreliable only because their initial responses to moral cases are often flawed. Elicitation methods which include scope for reconsideration allow people themselves to improve their judgments. However, our empirical research also highlights that people's pursuit of consistency is limited, as a sizable share of inconsistencies, even openly recognised ones, remained unresolved.

Thursday 2nd July 17:00 — Measurement & Models (Voorkamer)

Vuk Kolarevic: *Does Formalising Psychological Theories Advance Scientific Understanding?*

Building on a case study of a formalised theory in psychology, this paper advances a novel account of the evaluation of degrees of understanding. Central to the account is a new evaluative dimension I call a theory's grip, which concerns how tightly a given characterisation of a phenomenon fits with the relevant explanatory theory.

Overview: Psychology is often said to face a theory crisis, commonly attributed to its reliance on verbal theories rather than formal (mathematical) models. This has motivated recent calls to formalise psychological theories as a promising direction for advancing the field. However, critics argue that lack of formalisation is not a problem in itself: many successful scientific theories are non-formal, and premature formalisation can be counterproductive. This raises a central question:

under what conditions can formal modelling in psychology genuinely be of epistemic benefit? A promising answer appeals to scientific understanding. In fact, advocates of formalisation in psychology often claim that formal models enhance our understanding of the target phenomena as well as the relevant theories. However, within these debates, there has been little engagement with the extensive philosophical literature on scientific understanding that has developed over the past two decades. As a result, the notion of understanding invoked in these discus- sions remains under-theoried, and we are left without a clear account of how and why formal modelling is supposed to improve it. At the same time, philosophical accounts of scientific understanding have almost without exception already been focused on understanding brought about by formal theories (or models). Given this, one may assume that formal modelling is (perhaps tacitly) taken as a necessary condition for achieving under-standing within the philosophical literature. However, taking Henk de Regt's (2017) seminal account as a backdrop, neither formal modelling nor mathematically described theories are a requirement for achieving genuine scientific understanding.

While there are good reasons for this view, not least of which is that we should aim for our accounts of understanding to be compatible with the scientific practice of most of psychology (which in fact operates primarily with verbal theories), I maintain that the ramifications of such a view have not been sufficiently examined. Namely, de Regt's framework struggles to accommodate the intuition widely shared within the formalisation movement — that formalising psychological theories often advances the quality, or degree, of understanding that the given theories provide. On de Regt's view, the degree of understanding a theory T provides depends solely on the degree of T's intelligibility. However, formalised theories in psychology are often less intelligible to practitioners, and this is widely referenced as the central reason why formal modelling still hast caught on in the field. The problem is not avoided by appeal to other proposed evaluative dimensions of understanding, such as representational accuracy (or depth), since most formalised models of psychological processes are explicitly provisional and not accurate in the relevant sense.

The Aim of the Paper: The present paper has two central aims: 1. To show why the potential noetic value of formal modelling in psychology cannot be straightforwardly accounted for within existing accounts. 2. To propose a novel evaluative dimension of understanding, sufficient to explain how and why formalising psychological theories may in fact advance scientific understanding of the relevant phenomena.

My Proposal: In the present paper, I argue that due to the overwhelming focus on already formalised theories within the research on scientific understanding in philosophy, shedding light on the potential noetic value of formalisation efforts in psychology is not at all straightforward. Thus, I take a reverse route: I argue that practices and insights associated with the formalisation movement in psychology could instead productively inform our philosophical accounts of scientific understanding. In a nutshell, drawing on a case example of a formalised theory in psychology, I introduce a novel evaluative criterion of scientific understanding I call a theory's Grip.

The case example I discuss is the influential regulatory resource theory in psychology. Recently, van Dongen et al. (2025), forcefully addressed a crucial flaw of the theory that is characteristic of much of the psychological science — namely, that it is not clear what exactly the theory is supposed to predict. Since it leaves many crucial experimental assumptions underspecified, the authors remark that the theory's predictions "depend on how an individual researcher fills in the gaps and vagaries with their unstated personal assumptions and mental simulations". This is evidenced by the fact that, during the most recent multi-lab replication attempt of the resource theory, researchers had to consult the original authors, as key details needed to experimentally

test the theory were unclear from the original papers themselves. Based on this case example and van Dongen and coworkers' formal model of the regulatory resource theory,

I argue that formalising a verbal theory may contribute to the following three mutually reinforcing factors: (i) Understanding the theory (UT): gaining a clearer understanding of what the theory precisely implies. (ii) Antecedent understanding of the phenomenon (AUP): gaining a clearer understanding of what the phenomenon to be explained is in the first place. (iii) Grip: gaining a clearer grasp of how the theory constrains theoretical expectations about the given phenomenon.

On the view I propose, the degree of understanding a theory affords depends jointly on its degree of intelligibility and on the degree of grip a researcher has on the target phenomenon on the basis of that theory. The central advantage of this approach is that formalisation is treated as a tool rather than as a necessary condition for understanding. The account developed in this paper does justice to ongoing efforts to formalise theories in psychology and represents a fruitful addition to de Regt's original pragmatic and anti-realist-friendly account of scientific understanding.

Charlotte Constanze Poller: *Between Thickness and Comparability: Measurement Drivers in Global Well-Being Metrics*

Global measurements of well-being shape how individuals, institutions, and governments understand and pursue the "good life". The World Happiness Report (WHR) is one of the most influential: it publishes annual cross-national rankings based on survey data and links "happiness" to social and economic indicators, positioning itself as an empirical basis for public-policy debate (Helliwell et al. 2025). The WHR thus provides a paradigmatic case for examining how complex, value-laden concepts such as well-being are rendered measurable globally. To analyze the WHR's measurement strategy, this paper draws on Basso and Alexandrova's (2025) framework of measurement "drivers," understood as competing epistemic, ethical, pragmatic, and metrological considerations that guide design choices and inevitably force trade-offs. I argue that in the WHR pragmatic and metrological drivers, especially global comparability and communicability, are systematically privileged over epistemic and ethical drivers, yielding a deliberately "thin" operationalization of well-being. This thinness is particularly visible in the WHR's core measurement of well-being, which relies on a single life-evaluation item asking respondents to place their current life on a 0–10 ladder. Although this measurement is presented as neutral and universally applicable, its normative underdetermination leaves unspecified how people arrive at their evaluations and which aspects of life they take to matter for their well-being. While the WHR supplements its core life-evaluation measure with six explanatory factors, these function as predictors of reported well-being rather than as its constitutive dimensions, and thus do not resolve the underlying normative indeterminacy. T

his design choice is not without problems. Recent work in the philosophy of social science emphasizes that policy-relevant indicators do not merely inform decisions but also shape how those decisions can be publicly justified and criticized. In this vein, Thoma (2024) argues that, under conditions of persistent value disagreement, policy-relevant indicators should not restrict the public articulation and contestation of competing value commitments. This requirement becomes especially demanding in the case of well-being. As Alexandrova (2017) emphasizes, well-being is a thick, context-dependent concept: different societies endorse different considerations as constitutive of a good life. Since such considerations vary across contexts, global well-being measurement confronts persistent disagreement over which aspects are constitutive of well-being. This generates a conceptual dilemma: adequate measurement requires some degree of thickness in order to avoid arbitrariness and to legitimize policy guidance, yet substantial value-disagreement undermines the abstraction and standardization on which global comparison depends. How thick, then, can global well-being measures be while remaining

comparable? I argue that the WHR resolves this tension by prioritizing pragmatic and metrological drivers over epistemic and ethical ones, effectively standardizing away substantively relevant differences in what well-being is taken to be. However, this solution comes at the cost of normative adequacy.

Drawing on Thoma's (2024) argument that policy-relevant indicators should not foreclose the articulation of plural value commitments, I defend a minimum-thickness requirement for global well-being measurement. The core idea is not to replace thin global metrics with a comprehensive theory of well-being, but to rebalance competing measurement drivers so as to make at least some normative commitments explicit and contestable. I suggest that this can be pursued by modestly enriching the WHR's life-evaluation measure with a small set of explicitly normative dimensions, and by presenting these dimensions in a pluralistic, dashboard-like format.

Together, these moves shift priority toward epistemic and ethical drivers, at the expense of maximal comparability and communicative simplicity, and thereby making explicit the normative trade-offs involved in global well-being measurement.

Eyup Engin Kucuk, Maeve Burwell and Ömer Dağlar Tanrikulu: *What do Bayesian models explain? An empirical analysis of explanatory stances in Bayesian cognitive science*

Bayesian models have played a central role in cognitive science for over three decades and provided a widely used framework for studying perception, learning, and reasoning. They are sometimes presented as capturing the universal structure of cognition (Griffiths et al., 2011) or as enabling reverse engineering of the mind (Griffiths, Chater, & Tenenbaum, 2024). Despite their success, debate persists about what Bayesian models explain about human mind/brain. Some argue that Bayesian models are purely computational and carry no ontological commitments (e.g., Griffiths et al., 2012), while critics note that they are often treated as implying that the mind literally performs Bayesian inference (e.g., Bowers & Davis, 2012). Others claim that the evidence is insufficient for probabilistic mental representations and favor an instrumental interpretation (e.g., Block, 2018), while proponents argue that the success of Bayesian models supports their psychological reality (e.g., Rescorla, 2025).

Much of the debate is a dispute about how Bayesian models are used in practice, with each side attributing a stance to "Bayesian researchers" and the other side rejecting that attribution. This is illustrated by the commentaries on Jones and Love's (2011) Behavioral and Brain Sciences article. Although leading scholars disagree on many points, yet converge on a single conclusion: there is a widespread confusion about what explanatory commitments Bayesian models are meant to express. While prior work offers important theoretical analyses, there remains no standardized and scalable method for empirically characterizing explanatory stance in the literature at scale. Our goal in this work is to analyze the language in academic articles to uncover the implicit assumptions authors make when using Bayesian concepts.

Here we present a large-scale empirical analysis of explanatory stances in Bayesian cognitive science using a theory-driven annotation framework. We introduce a codebook for identifying "assumption-bearing quotes": sentences that make claims about Bayesian modeling/inference as explanations of behavioral, cognitive, computational, or neural processes (Table 1). The codebook maps these claims onto two continuous and mirroring scales: "realism" and "instrumentalism". Realism treats Bayesian models as describing the actual mechanism of the human mind, while instrumentalism considers them as useful tools without ontological commitments. We adopt this framework because it reflects the current debates in Bayesian

cognitive science, subsumes previous theoretical distinctions in the literature, and aligns with the broader realism-instrumentalism tradition in the philosophy of science.

To validate this framework prior to the main analysis, 250 quotes from 15 articles were extracted and hand-annotated by an expert. The data were split into training (80%) and test (20%) sets, and three different large language models (LLMs) were prompted with the codebook and training examples. LLM–human agreement on categorical assignments was moderate for Gemini (κ=.45-realism; κ=.48-instrumentalism), OpenAI (κ=.60-realism; κ=.59-instrumentalism), and Claude (κ=.54-realism; κ=.52-instrumentalism). For continuous scores, agreement with human ratings was high for Claude (ICC(2,1)=.90-realism & instrumentalism; ICC(2,k)=.95) and moderate-to-high for OpenAI (ICC(2,1)=.70-realism & instrumentalism; ICC(2,k)=.82). Overall, these results indicate that the models could apply the codebook reliably enough to support the subsequent analyses.

Using this framework, we analyzed 6,941 assumption-bearing quotes from 211 peer-reviewed cognitive science articles. We built the corpus by seeding 50 APA database results from the relevant Bayesian modeling keyword search and screening abstracts for substantive engagement with Bayesian modeling (empirical applications, computational modeling, or theoretical discussion; not mere statistical use). We then expanded the set using ResearchRabbit's citation-network map with additional abstract screening. Across the three models, annotation agreement was moderate for categorical subcategories (Fleiss' κ=.48-for realism; κ=.47-for instrumentalism) and high for continuous scores (ICC(2,1)=.75-for realism; ICC(2,1)=.74-for instrumentalism; ICC(2,k)≈.90-across both scales). We then averaged the three models' continuous realism and instrumentalism scores per quote and used these averaged scores in downstream analyses.

These analyses presented three important results. First, intraclass correlation analyses revealed substantial within-paper heterogeneity: between-article variance (SD = 14.7) was smaller than within-article variance (SD = 20.8), with only 33.5% of variance attributable to between-paper differences (ICC = .34) (Figure 1). This pattern indicates that explanatory commitments vary substantially within articles: individual papers often contain conflicting realist and instrumentalist assumptions in their language.

Second, explanatory stance differed systematically by domain. Here, domain refers to the level of process a paper primarily targets: lower-level perceptual and motor processes versus higher-level cognitive processes. We fit a linear mixed-effects model predicting quote-level realism from domain level (high vs. low), article type (computational, experimental, and theoretical), and publication year. The model included random intercepts for articles to account for multiple quotes per paper. Articles focused on lower-level were more realist than those focused on higher-level ($\beta = 7.60$, SE = 2.03, $t = 3.75$). Including domain level improved model fit ($\chi^2(1) = 13.63$, $p < .001$) (Figure 2). This effect remained while controlling for article type and publication year ($\beta = 9.57$, SE = 2.05, $t = 4.68$). Article type did not improve the model fit ($\chi^2(2) = 5.73$, $p = .057$), as publication year did not reliably predict paper's stance (realism: $\chi^2(1)=3.56$, $p=.059$; instrumentalism: $\chi^2(1)=3.44$, $p=.064$).

Third, averaging quote scores within articles and comparing realism/instrumentalism scores enabled us to classify 160 (75.8%) articles as instrumentalist-leaning, 39 (18.5%) as realist-leaning, and 12 (5.7%) as mixed/ambivalent (weighted by quote volume: 5,279 (76%) instrumentalist-leaning quotes, 1,332 (19.2%) realist-leaning quotes, and 330 (4.8%) mixed/ambivalent quotes). This result suggests that cognitive science literature shows an overall instrumentalist tendency, both at the article level and at the level of individual quoted claims.

By turning a long-standing theoretical dispute into a measurable empirical target, this work provides a quantitative map of how Bayesian explanations are framed across cognitive science

and offers a scalable method for discourse/stance analysis of explanatory commitments using LLM-assisted annotation. We have achieved three main results through our work: (1) individual research articles frequently mix realist and instrumentalist language, (2) explanatory stance shifts between high- and low-level cognition fields, and (3) there is an instrumentalist tendency in the literature overall. These results provide empirical support for the previous theoretical discussions regarding the "confusion" about the role of Bayesian models in explaining human mind and also the tension between explicit and implicit assumption within research articles using Bayesian modeling.

Yara Daamen, Vlasta Sikimić and Daniël Lakens: *The Intellectual Justice Scale: Development and Validation of a Self-Report Measure*

Intellectual justice, the disposition to engage fairly with knowledge and those who produce it, has received growing attention in philosophy and education, yet no validated self-report measure exists. This research describes the development and validation of the Intellectual Justice Scale (IJS), a 9-item measure with three subscales (Testimonial Justice, Hermeneutical Justice, and Intellectual Action) grounded in the framework of epistemic injustice (Fricker, 2007). Being able to measure individual differences in intellectual justice is an important step toward studying whether this virtue varies across people and whether it can be cultivated through education or targeted interventions.

Scale development followed established guidelines (Carpenter, 2017; Hinkin, 2005; Morgado et al., 2017). An initial pool of 66 items was developed based on Fricker's (2007) theoretical framework, refined through expert input from philosophers specialising in intellectual justice, and evaluated through several rounds of iterative review. A pilot study (N = 100) further reduced the pool using descriptive item statistics and qualitative judgment, resulting in 22 items carried forward for the main studies. In Study 1 (N = 404), exploratory factor analysis supported a three-factor structure. Item selection was conducted independently by two researchers, one using statistical criteria and one using conceptual criteria, and both selections converged on the same nine items, three per subscale. The 9-item scale showed good internal consistency ($\alpha = .84$, $\omega = .87$), with subscale reliabilities ranging from $\alpha = .72$ to .74. In Study 2 (N = 401), confirmatory factor analysis was used to evaluate the factor structure in an independent sample. A higher-order model, in which Testimonial Justice, Hermeneutical Justice, and Intellectual Action are the dimensions of a single overarching intellectual justice construct, showed good fit (robust CFI = .964, robust TLI = .946, robust RMSEA = .068, SRMR = .037). All factor loadings were significant ($p < .001$). The scale again showed good reliability ($\alpha = .82$, $\omega = .85$).

To assess construct validity, participants additionally completed the Intellectual Humility Scale (Leary et al., 2017), the HEXACO-60 personality inventory (Ashton & Lee, 2018), the Social Justice Scale (Torres-Harding et al., 2011), and the Justice Sensitivity Short Scales-8 (Groskurth et al., 2023). As expected, the IJS showed positive associations with all validity measures. The strongest associations were found with social justice attitudes ($r = .82$) and intellectual humility ($r = .67$), reflecting meaningful conceptual overlap with broader justice orientations and related intellectual virtues. Associations with personality traits were moderate (Openness to Experience: $r = .50$; Agreeableness: $r = .49$) to weak (Honesty-Humility: $r = .29$; Conscientiousness: $r = .21$), and justice sensitivity showed a similarly weak association ($r = .27$), suggesting that intellectual justice is not simply a reflection of broader personality dimensions. Discriminant validity analyses indicated that the IJS remained empirically distinct from all included measures, with a theoretically interpretable marginal overlap with social justice attitudes.

Taken together, these findings offer the first validated self-report measure of intellectual justice attitudes. The three-subscale structure provides empirical support for Fricker's theoretical

distinction between testimonial and hermeneutical injustice, while the Intellectual Action subscale extends this framework by capturing active epistemic engagement. The IJS has potential applications in research on epistemic fairness, educational interventions aimed at cultivating intellectual virtue, and motivational attitudes related to critical thinking in AI.

Thursday 2nd July 17:00 — Mindreading (Bovenkamer)

Martin Doherty and Catherine Sayer: *Preschool development of the concept of line-of-sight as evidence for developmentally distinct systems of gaze processing*

Infants follow others' gaze direction from at least the second year of life yet three-year-olds cannot judge where someone else is looking (Doherty et al., 2009). A potential explanation for this disjoint is that gaze following and gaze judgement depend on developmentally distinct psychological systems (Doherty 2011). One orients the child's attention to objects others attend to, may be evolutionarily old, and hypothetically only outputs the object of attention to the general cognitive system. The other is about the relation between the agent and the object as such. It forms an important part of thinking about others' mental states and hypothetically develops as part of general theory of mind development. Distinguishing gaze following from gaze judgement poses methodological challenges. Seeing eyes directed towards a target automatically cues attention to the target, which is then more likely to be chosen regardless of whether a participant can judge the gaze relation. As an alternative we use the concept of line-of-sight. This is the understanding that there must be a straight uninterrupted line between an agent and an object for the agent to see that object. This is a fact about the gaze relation; understanding this fact indicates the ability to think about this relation.

This talk will present data on the development of this concept. We will briefly review two studies (total N = 200) examining children's judgement of line-of-sight. We showed children small scale models in which a target was occluded by either a curved or a sharp corner. Young preschoolers judged another could see the target. Three-year-olds continued to do so even after looking themselves from the other's vantage point. Effects were greater for curved corners, possibly indicating they had learned about sharp corners from experience. Performance was significantly associated with the ability to judge eye direction. This supports the claim that a concept of line-of-sight develops between 3- to 4-years. A strong prima facie objection to this conclusion is children's facility with other perspective taking tasks. Even 2-year-olds can hide objects by placing them behind occluding objects, or in other words by placing them where there is no line-of-sight. However, they are unable to hide objects by placing occluders in front of them, suggesting their success is not via manipulating line-of-sight (Flavell et al., 1978). Instead we argue that early success is via general sensitivity to what others are likely to engage with (e.g., things in front of you, with no occluding barriers). Engagement persists once established, and placing small occluders does not subsequently affect it.

We briefly review two further studies looking at children's hiding ability (total N = 144). Manipulating whether engagement has been established by blindfolding the agent before introduction of the target object improves ability to interpose occluding objects. This supports the claim that younger children approach the task by judging general involvement rather than visual perspective. Performance on the original 'move-screen' task associates cross-sectionally with gaze judgement and theory of mind tasks. Additionally there is modest evidence of longitudinal links between early gaze judgement and later theory of mind performance, providing limited support for gaze understanding as an early manifestation of theory of mind abilities. Overall we consider there is a good empirical case that line-of-sight understanding develops in the preschool period. It is a clear conceptual consequence of understanding the gaze relation between agents and objects. This supports the claim that gaze processing is based on distinct systems, one early but limited, and one developing in preschool. Potential links to theory of mind development

suggest that the later gaze processing system and theory of mind are developmentally related. The causal nature of this development is the focus of ongoing work.

Calum Sims: *Submetacognition: I am not really reading my mind*

In the debate between proponents of associative and non-associative explanations of animal behaviour (i.e. Smith 2009, Beran 2019, Crystal 2009, Crystal 2019), Heyes' (2014) 'submentalising' account proposed that complex social behaviours could be explained without invoking the ability to represent the mental states of other agents. Heyes argued that a subordinate ape's ability to mislead a dominant ape into pursuing an inferior food source could be explained without invoking that subordinate ape's ability to know what the dominant was thinking. Rather than needing to represent things like, 'the dominant ape believes that I will lead him to the best food-source' or 'the dominant ape is not aware of the superior food source', the subordinate ape can simply track behavioural features of the environment - most significantly, the gaze of the dominant and his relation to the food-source - in order to form the sorts of predictions that allow him to guide his action flexibly enough to engage in complex behaviours like deceit. This submentalising interpretation, also dubbed 'behaviour-reading', shows how complex and apparently sociocognitive behaviours can get off the ground without complex representational abilities. This is significant because such abilities are taken to explain the differential flexibility of human and nonhuman animals' behaviour. The submentalising interpretation opens the door to the view that complex (i.e. flexible, adaptive) behaviours can be explained without invoking significant cognitive complexity, and in this spirit goes some way to closing the gap between accounts of human and animal behaviour.

The purpose of this paper is to extend this line of reasoning from mindreading to metacognition, developing a 'submetacognising' or behaviour-reading account of metacognition. The basic principle of this account is simple: when guiding one's own action flexibly, it is rarely, perhaps never, necessary to metarepresent one's mental states to oneself; rather, representing or otherwise accessing one's own behavioural states is sufficient. This account suggests that one's own mental state is inferred by the agent from his or her own behaviour - rather than 'looking within' and accessing some mental state that tells us what we're thinking, we construct our thoughts on-the-fly by inner observation of these behavioural states.

This paper explores a range of human and animal experiments to argue that submetacognising interpretations are powerful explanatory tools in both human and nonhuman animal cognition. The paper focuses on a range of experimental set-ups from comparative cognition (particularly, 2AFC, mirror recognition, perspective-taking and planning) to argue that submetacognitive interpretations can explain the flexibility of both human and nonhuman animal behaviours. I then explore the implications of this for the view that metacognition explains human uniqueness, sketching a non-cognitivist account of metacognition that distinguishes humans from nonhuman animals in terms of the changes we make to our ecological niche.

William Angkasa: *Affective MIND Script: A Situated Cognitive Architecture for Everyday Social Interaction*

Understanding how humans navigate social interactions is central to social cognition research. Mindreading, within traditional folk psychology, emphasizes inferring mental states as critical for predicting and coordinating behavior. Competing frameworks (particularly those grounded in situated cognition and affectivity) critique this mindreading-centric approach for overestimating our reliance on high-cost inferential processes. Studies in behavioral ecology suggest that humans reduce cognitive demands by relying on heuristics and context-sensitive routines, making

social scripts a compelling alternative. Recent work on scripts in social cognition has begun to integrate cognitive, structural, and normative dimensions, yet these strands often develop in parallel. In particular, the role of affectivity as the mechanism that shapes how scripts are embodied, stabilized, and recalibrated in practice has not been systematically theorized in this context.

This paper proposes the Affective MIND Script framework, which reconceptualizes scripts as Modular, Internalized, Negotiable, and Distributed entities. It treats affectivity as the modulatory engine that structures and regulates scripts synchronically and diachronically as they unfold across contexts. By characterizing scripts as low-cost, pre-learned heuristics embedded in dynamic interaction, the framework offers an ecologically grounded alternative to mindreading-centric models of social cognition. On this view, mindreading and mindshaping are secondary, higher-cost strategies typically recruited when script-based coordination falters. The account clarifies how affectively charged scripts enable individuals to seamlessly participate, negotiate, and adapt in everyday social interaction.

Ayse Payir, Kathleen Corriveau and Paul L. Harris: *Believing without seeing: Children Expect Others to Share their Beliefs about Invisible Entities*

Children learn about invisible scientific and religious entities via testimony, relying largely on the consensus surrounding these entities, rather than on first-hand experience (Ma et al., 2024). However, studies have mainly examined children's claims about their own beliefs regarding high consensus entities, such as germs and God. How do children judge controversial entities, such as evolution? How do they evaluate others' beliefs across both immediate and broader social contexts? And to what degree does this evaluation demonstrate children's awareness of belief diversity?

To explore these questions, we presented 116 US children (6- to 12-year-olds, M = 8.36, SD = 1.74) with 12 entities that vary in the community consensus surrounding them. Based on similar research conducted with adults (i.e., Shtulman, 2014), children judged: 1) if an entity is real or not, 2) how confident they are in this belief (5-very sure to 1-not sure at all), 3) how many of their family and friends, and 4) how many Americans hold the same belief (5-all of them to 1-none of them). We created a 10-point belief score (10-Real, very sure, to 1-Not real, very sure) by merging the reality judgments with confidence ratings. We also created a 10-point consensus score (10-Real, all of them agree, to 1-Not real, all of them agree), separately for friends and family, and for Americans, by merging reality judgments with consensus estimations. Children expressed a strong belief in scientific entities regardless of religious background, ($p = .65$; Figure 1), and they expected their family and friends (Figure 2, left), as well other Americans (Figure 2, right), to share this belief ($ps \geq .46$). An exception to this pattern was evolution, for which children with a non-religious background expressed a stronger belief, ($p = 0.005$), and attributed greater consensus to both their family and friends, ($p = 0.003$), and to Americans, ($p = .017$), compared to children with a religious background. Background also influenced belief in high-consensus religious entities; children with a religious background not only expressed a stronger belief in these entities ($p < .001$), but also expected both family and friends, ($p < .001$), and all Americans ($p < .001$), to join them in their belief.

These findings show that religiosity is not associated with a lower belief in scientific entities among US children (except for evolution). Belief in religious entities, however, is largely determined by religious background, and can vary sharply across the religious and non-religious. An important novel result is that regardless of religious background, children expect not just members of their immediate context but also members of the broader social context to agree with them. Although

religious and secular children may be right about members of their immediate context, their divergent claims about the broader American context cannot both be right. Future research should explore whether this misperception of the broader context holds in other political and cultural contexts.

Friday 3rd July 09:00 — Keynote (Kerkzaal)

Tadeusz Wiesław Zawidzki: *What Is Mindshaping?*

Mindshaping is an alternative to orthodox accounts of human competence at social coordination. Human beings are extreme biological outliers in this domain: no other primate species approaches our capacities at coordination on complex, long-term, large-scale cooperative projects with unfamiliar others. Most of our unique capacities for technological innovation, threat mitigation, and resource extraction depend on this; so, it is no exaggeration that explaining uniquely human capacities for coordination is one of the most significant projects of the biological and social sciences.

On the orthodox account of this competence, it relies on our unique talents for "mindreading": we are much better than other primates at reliably inferring each other's thoughts, and this explains why we are so much better at social coordination. However, there are deep problems with this admittedly compelling picture. Most significantly, given the holism of human thought, it is not clear how we manage to reliably track each other's thinking. Human behavior depends on indefinitely large, whole networks of cognitive and conative states. For this reason, any finite set of observable behaviors or situational factors is compatible with an infinite set of possible thoughts. E.g., the same stimulus may cause fear in one person and indifference in another, depending on differences in what each knows about the stimulus. It is not clear how any of the classical theories of mindreading, e.g., so-called "theory-theory" or "simulation theory" or other alternatives, can solve or mitigate the holism problem.

According to the mindshaping hypothesis, the holism problem is mitigated through mechanisms and practices, like fine-grained imitation, pedagogy, norm cognition, and narrative self-constitution, aimed at shaping human minds to be more alike, and hence, more easily interpretable and predictable. When interpreting others, we do not need to consider most possibilities compatible with observable behaviors or situational factors, because most of the people we need to interpret and predict have been shaped (as we have) to find familiar facts relevant and familiar outcomes preferable. Such shaping begins at birth, but continues throughout life, crucially involving self-regulation aimed at making our own behavior fit social roles that make us easily interpretable relative to our communities.

In this talk, I review some historical antecedents of mindshaping, including Aristotelian "second nature" (McDowell 1994), Kierkegaardian "pretence" (Lear 2011), various notions of "ideology" (including Nietzsche's, Marx's, Althusser's, Deleuze's, Foucault's), Heidegger's notion of "das Man" (Dreyfus 1995), Goffman's "presentation of the self in everyday life" (2021), and Vygotskyan approaches to developmental psychology (Vygotsky 2012, Tomasello 2019). I then consider some classic discussions of the orthodox view, e.g., Lewis's analysis of convention (1969), arguing that they fail to account for coordination, due to the holism problem. Next, I ground mindshaping in a philosophical framework according to which social cognition is a species of norm cognition (Brandom 1994) and review some relevant empirical findings. Finally, I address the main challenge to the mindshaping hypothesis - bootstrapping: how can we shape each other to meet each other's expectations without prior, reliable mindreading?

Friday 3rd July 10:45 — Plenary Symposium (Kerkzaal)

Ian Apperly, Víctor Fernández Castro & Ildikó Király: *Mindshaping: Development, Diversity, and Individual Differences*

Ian Apperly: *A Mindshaping Perspective on Diversity and Individual differences in Mindreading*

One objective of the mindshaping framework is to challenge claims that mindreading is the principal explanation of human social abilities. A second objective is to explain how potentially intractable social reasoning is made tractable. I will argue that this helps us think about a significant neglected puzzle: Why do adults who clearly have the full suite of mental concepts continue to vary in their mindreading abilities? Mindshaping implies that the ability to use mental concepts for mindreading depends upon socially constructed expectations, schemas, and norms. I subject myself to these in order to appear rational and reasonable, and you can apply them to guide inferences about my thoughts and feelings. It is, however, likely that our exposure to these expectations, schemas, and norms will vary. You should therefore find it easier to mindread me if we are more similar, and you will be a better mindreader overall if your mindreading has the flexibility to cope with varying degrees of your similarity to other people. I will present evidence in favour of both predictions from a new mindreading task based upon “crowdsourced” stimuli that capture ecologically valid variability in mindreading explanations, and a diverse sample of 2500 adolescents and adults. I conclude that mindreading involves the flexible application of mental concepts. This is enabled by mindshaping, and it varies because we are not shaped equally.

Víctor Fernández Castro (University of Granada): *The Neurodiversity Paradox: a mindshaping solution*

In recent decades, the neurodiversity movement has sought to create conceptual space for recognizing atypical or divergent cognitive profiles as potentially constitutive of personal and social identity. Inspired by biodiversity, this movement holds that cognitive diversity is a natural and valuable phenomenon for collective functioning. Its activists reject both the default pathologization of natural human variants (such as autism or schizophrenia) by institutional psychiatry and the default normalization of their conditions, which risks downplaying their disabling character in certain contexts while still insisting they can be integrated into personal or social identity.

Following Walker (2021), the movement can be defined by three claims: neurodiversity is a natural and valuable form of human diversity; the idea of one normal or healthy type of brain or mind is a culturally constructed fiction; and the social dynamics manifesting in regard to neurodiversity are similar to those of other forms of human diversity, including power inequalities and creative potential. However, this definition reveals an important tension between two claims: cognitive diversity as a natural phenomenon (biologically preconfigured) and openness to contestation (the categories through which we conceptualize this phenomenon are open to social contestation).

To resolve this tension, I draw on the mindshaping view (Zawidzki 2013, 2021). According to this view, when we ascribe or self-ascribe mental concepts (such as beliefs or desires) or socially recognizable identity concepts (such as child, spouse, or student), we are tacitly triggering, justifying, and specifying relevant norms of how to behave, cognize, and feel. These discursive conceptualizations help shape behavioral dispositions and expectations that facilitate coordination on cooperative projects. These projective features have two dimensions: a backward-looking evaluation, in which a person's behavior is justified by the identity judgment or mental state ascription, and a forward-looking prescription, in which the person is X, therefore

they should act in certain ways from now on. These dimensions manifest in various uses of ascriptions: justificatory and exculpatory purposes, pedagogical purposes, showing inconsistencies, or declaring commitments.

This view of discursive conceptualization helps address the neurodiversity movement's tension. At the sub-personal level, the mindshaping view is compatible with variable biological preconfigurations of traits and dispositions. At the personal level, some of these traits are partially conceptualized using terms like "autistic" or "ADHDer," recognizing that these terms have projective and regulative functions. The fact that this projective character has a normative dimension specified by commitments means those commitments can be contestable. As McGeer (2019) puts it, “mindshaping is inescapable, social injustice is not”. That identity and mental concepts are associated with regulative practices does not imply the norms themselves cannot be questioned. What counts as behaving well or badly under a concept is itself contested and contestable terrain. Thus, we can view the neurodivergent struggle as a struggle over the normative expectations associated with identity and mental concepts. It disputes that such expectations must be naturalized as part of biological preconfiguration, especially those that assume deficit by default. This does not deny biological cognitive styles; it simply means our social niche should not assume that all social expectations created by mindshaping dynamics are necessarily based on such biological dispositions. What is being contested is not merely a matter of disputed scientific facts. Rather, the neurodiversity movement is contesting the very normative content of the concepts we use, concepts that have fundamental consequences for how we navigate our social environment and, more importantly, how we perceive ourselves.

References

McGeer, V. (2019). Mindshaping is inescapable, social injustice is not: Reflections on Haslanger's critical social theory. *Australasian Philosophical Review*, 3(1), 48–59.
Walker, N. (2021). *Neuroqueer heresies: Notes on the neurodiversity paradigm, autistic empowerment, and postnormal possibilities*. Autonomous Press.
Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.
Zawidzki, T. W. (2021). Suffering and mindfulness: A Neo-Darwinian perspective. *Journal of Buddhist Philosophy*, 3(1), 36–48.

Ildikó Király: *Mindshaping in the context of development*

Mindshaping is based on the premise that some cognitive tools —such as (over)imitation, natural pedagogy, and the imitation of fictional agents—enable people to cooperate effectively and successfully without the application of mindreading (theory of mind), namely without having to pay attention to the thoughts or mental states of others. These competencies are available to children at a very early age, but the mechanisms underpinning them are controversial: is mindreading necessary for them to be functional and flexible? The mindshaping model emphasizes that these tools can and should support behavioral coordination and cooperation without attributing intentions and beliefs to others. In this presentation, I will present empirical evidence on the characteristics and central mechanisms of the above competencies in early childhood. The focal question is whether the tools of imitation, pedagogy, and understanding fictional characters are suitable for achieving subtle social cooperation or not, on their own, without simple forms of theory of mind.

Friday 3rd July 14:30 — Symposium: Inner Speech (Kerkzaal)

The nature of inner speech is a growth field in philosophy, psychology, neuroscience and linguistics. The large number of publications on inner speech within the last ten years in all of these areas attests to its importance. Inner speech raises fundamental philosophical and empirical questions. Is there unconscious as well as conscious inner speech? Is inner speech literally a kind of speech, or is it only an imaginative representation of speech? What properties does it share with overt speech? Is there a difference between the speech acts possible in inner speech and those possible in overt speech? What is the function of inner speech? If it is merely a supplement to conceptual thought, what does it contribute? If it is itself a medium of conceptual thought, is it only one among others, or is it the sole medium of all conceptual thought? What are the faculties of mind that are responsible for the production of inner speech? What might be the roles of imagination, motor systems, and memory?

The symposium here proposed will bring together four accomplished inner speech researchers, whose presentations will illustrate four kinds of questions that can be asked about inner speech. Bo Yao, a neuroscientist from Lancaster University, will explain how we can test the proposition that spontaneous inner speech (as opposed to experimenter-provided verbal cues) facilitates perceptual object recognition. Dr. Yao takes this as support for a conception of inner speech as facilitating predictive processing. This talk will illustrate the study of the psychophysical effects of inner speech. Jutta Mueller, a cognitive scientist from the University of Vienna, will present data indicating that in people who frequently engage in conscious inner speech there is more cross-talk between the brain's language network and areas of the brain responsible for the integration of information, and there is more cross-talk between the brain's default-mode network and regions responsible for such things as semantic memory. This talk will illustrate the study of the function of inner speech in the economy of the brain. Nikola Kompa, a philosopher from the University of Osnabrück, will argue that the internalization of speech can account for the formation of beliefs of the kind that we think of as functioning to determine one's social identity. This talk will illustrate the relevance of the topic of inner speech for social philosophy. Finally, Christopher Gauker, a philosopher from the University of Salzburg, will argue that inner speech should be regarded as a product of a more general capacity that he labels constructive imagination. This talk will illustrate debates about the nature of inner speech as mental representation.

Bo Yao: *How does inner speech resolve visual uncertainty? Testing the precision modulation account*

Inner speech - the silent production of language in our minds - remains theoretically elusive despite decades of research. This talk introduces Linguistic Active Inference Theory (LAIT), which proposes that inner speech augments the brain's predictive processes by deploying linguistic categorical priors that constrain perceptual inference under uncertainty. Through language's efficiency, extendibility, and generativity, inner speech anchors complex sensory experiences in linguistic forms for rapid categorical inference and unpacks abstract goals into situated actions for motor control. I present an empirical paradigm designed to test LAIT's core prediction: that self-generated inner speech modulates inference under perceptual uncertainty by deploying high-precision linguistic priors. Critically, existing evidence for language's influence on perception derives primarily from externally provided verbal cues - whether self-generated inner speech produces comparable effects remains underexplored. In this task, participants viewed videos of

everyday objects that progressively unblur over 10 seconds via Gaussian blur reduction; they pressed a key and typed their response once they recognised each object. This dynamic paradigm creates a continuous uncertainty gradient, capturing both recognition thresholds and decision timing. Participants completed the task under two counterbalanced conditions: articulatory suppression (continuously verbalising "aluminium" to block inner speech) versus foot tapping (a motor control task allowing speech processes). Individual differences in verbal thinking are measured using the Internal Representations Questionnaire. LAIT predicts a specific signature: when inner speech is used to make an inference, linguistic priors enable early but potentially premature categorical commitments, producing more variable decision times. This variability arises because some linguistically-driven guesses succeed early while others prove incorrect, requiring additional time for re-inference and error correction. Articulatory suppression should eliminate this variable guessing, forcing conservative evidence accumulation and yielding slower but more uniform responses. Crucially, this effect should be pronounced only in individuals who habitually rely on verbal thinking; low-verbal thinkers are likely to show minimal condition differences as they may not rely on language to resolve perceptual uncertainty. Time-course analysis will test whether the interference effect concentrates during high-uncertainty phases (early in trials when objects remain more blurred), diminishing as sensory evidence becomes conclusive. Pilot data (N=24) suggest that articulatory suppression does alter decision patterns in ways consistent with the elimination of linguistically-mediated priors, with effects moderated by verbal thinking preferences. The preregistered study will employ Bayesian sequential testing to provide conclusive evidence for LAIT's precision modulation mechanism using this paradigm.

Jutta L. Mueller: *Phenomenology of inner speech explains interindividual differences in the brain's resting-state networks*

Humans experience their thoughts, feelings and perceptions in very different ways. This is particularly evident in the domain of inner speech. While some perceive an almost constantly active inner conversation including a voice with rich sensory characteristics, others report the complete absence of inner speech. The experienced forms and contexts of inner speech can be quantified using introspective methods such as questionnaires or experience sampling methods. Several studies linked interindividual differences captured by such measures to performance on cognitive tasks. While there is some evidence that individuals who experience more inner speech perform differently on higher and lower-level cognitive tasks, e.g. meta-cognitive judgements or object perception, it is not known whether and how such differences are reflected in the human mind and brain beyond specific task contexts. If the experience of inner speech is a robust personal trait that generalizes over many contexts, one should assume that it is deeply rooted in the functional architecture of the brain. Recent years have brought many advances in the understanding of connectivity in the functional networks of the brain at rest and during task performance. Resting-state functional magnetic resonance imaging (fMRI) has discovered the so-called default-mode network (DMN), which is anti-correlated to demanding cognitive tasks and has been assigned the role of internally oriented mentalization, while the brain's language network (LN) – which partially overlaps with the DMN – is the task-related functional network that is typically engaged during language processing. In this study, we aimed to establish whether people's habit of engaging in internal verbal reasoning was related to the spatial pattern of brain coactivation with the brain's LN and DMN, during resting state fMRI. For this, we correlated interindividual differences in the experience of inner speech, as measured by an introspective questionnaire (internal reasoning questionnaire; IRQ), with the extent and spatial configuration of two brain functional connectivity networks at rest. We correlated scores on the IRQ verbal factor with two aspects 4 of whole brain connectivity, that involving the DMN and that involving the LN. In participants reporting high levels of inner speech we found higher levels of cross-talk between

the LN and right hemispheric brain regions known to be relevant for executive functions, visual-perceptual processing, sensory and multimodal integration and abstract reasoning. Similarly, in participants reporting higher levels of inner speech we found higher levels of cross-talk between the DMN and bilateral and left-lateralized brain regions known to mediate, e.g., semantic and self-referential memory and thought as well as speech or action control. Jointly, these findings suggest that participants' introspective reports about their inner speech habits go along with marked differences in their brains' networks even when no particular task is performed. On the one hand, the findings can provide a neurobiological foundation to the experiential reports. On the other hand, they converge with previous reports that intrinsic brain functional connectivity patterns reflect individual differences in skill and or performance, in our case extending this literature in showing that it also correlates with the propensity to experience inner speech.

Nikola Kompa: *Belief formation and inner speech*

According to the traditional view, humans form beliefs in a reason-responsive manner. However, recent debates in political epistemology and cognitive science concerning motivated reasoning, conspiratorial beliefs, and tribal thinking paint a different picture. In response, the idea that there are two types of belief has emerged (Westra, E., Philosophical Perspectives 2023; Mayer, M, et al. Cognition 2026). On the one hand, there are epistemic beliefs, which are formed on the basis of reasons or evidence. On the other hand, there are socalled symbolic or identity-representing beliefs, which serve a different function, such as signalling one's social identity. The question of how symbolic beliefs are formed has received rather little attention thus far. In this talk, I will examine the potential role of inner speech in this process. To this end, I will briefly introduce the account of inner speech that I favor. According to this broadly Vygotskian account, inner speech is internalised social speech. Language is initially acquired as a means of social interaction, but over time it becomes internalised, turning into a cognitive tool. Importantly, children internalise not only a system of labels and rules, but also a set of socio-linguistic practices such as argumentation, dialogue, storytelling, and joint problem-solving (Kompa, N., Bloomsbury 2024). However, they do not simply learn to engage in these practices offline. They also internalise the normative expectations that govern these practices, together with certain content and assumptions. They internalise 'the voice' of their social group (an idea found in the work of G. H. Mead and, more recently, M. Tomasello). They adopt these assumptions not because they have been given reasons for their truth, but rather without critical examination. However, by repeating these assumptions in their inner speech, they may actually come to believe them to be true. There is empirical evidence suggesting that people tend to believe something is true if they hear it repeatedly — the so-called 'truth effect' (Hasher, L. et al. Journal of Verbal Learning & Verbal Behavior 1977). Moreover, while these assumptions may not initially be explicit or fully integrated with other beliefs, expressing them through 5 inner or outer speech can make them more explicit (Frankish, K., Cambridge University Press 2004) and lets them guide behaviour.

Christopher Gauker: *The imagination theory of inner speech*

What capacity of the mind is the source for inner speech? Many authors say that an episode of inner speech is a product of a forward model, stimulated when commands to speak are sent to the motor systems of speech, which becomes conscious when the commands are aborted. But it is not plausible that the products of forward models always become conscious when the initiating motor commands are aborted, and this account fails to account for the interactions between inner speech and episodes of visual and auditory imagery. An alternative answer rests on a conception of constructive imagination in general. We use visual constructive imagination on a daily basis

when we figure out how to put objects together, for instance in deciding which arm hole of an overcoat to stick one's arm through or in figuring out how to wrap a sandwich in plastic wrap. We use constructive imagination, both visual and auditory, to form mental imagery of objects and events that we have never perceived. We should expect that imagery is constructed in accordance with methods of construction that operate on the structure of the quasi-perceptual representations that constitute mental imagery. Mental imagery is often unconscious, but it can become conscious as way of providing candidates for attention. According to the imagination theory of inner speech, inner speech is a special case of constructive imagination in which the imagery constructed represents an episode of speech. The methods of construction will be the same as those by which acts of speech are constructed, with the difference that what is constructed is only a representation of speech and is not an act of speech. These methods will ensure that episodes of inner speech represent grammatically correct utterances of sentences and that other aspects of coherent, useful discourse are adhered to. Inner speech, so conceived, may serve as the medium for a kind of a distinctive kind of thought. Much inner speech may be unconscious. The auditory imagery of inner speech is the special case of inner speech in which the representation of speech becomes phenomenally conscious. The function of the consciousness of inner speech is to provide candidates for attention. A consequence of the imagination theory of inner speech is that inner speech is not itself speech but only a representation of a nonexistent episode of speech. We can nonetheless say that inner speech also indirectly represents that which the nonexistent episode of speech represents. This kind of double representation is familiar from sound recordings of speech, which represent the sound of the recorded speech, but also represent indirectly the same things that the speech that was recorded represented. Such double representation characterizes also AI-produced sound representations of speech that do not record any actual speech.

Friday 3rd July 14:30 — Perception (Grote zolder)

Quentin Coudray & Assaf Weksler: *Grouping and the Metaphysics of High-Level Perceptual Experience*

High-levelists hold that perceptual experiences of high-level properties (such as natural or artifactual kinds) have their own phenomenal character (Bayne, 2009; Burnston, 2023; Cavedon-Taylor, 2021; Fish, 2013; Nanay, 2011; Ransom, 2020; Siegel, 2011; Stokes, 2018). There is, on this view, something it is like to visually experience a tulip, over and above the experience of its shape, color, texture, and other low-level properties.

Presently, there are two broad approaches to the metaphysics of perceptual experience in general. According to the 'what'-approach, phenomenal character is fully constituted by *what* objects and properties are experienced. Paradigm examples include strong versions of naïve realism (Campbell, 2006; Fish, 2009; Sethi, 2025), and of representationalism (Dretske, 1995; Pautz, 2021; Tye, 1995). According to the 'how'-approach, phenomenal character is constituted, at least in part, by *how* objects and properties are experienced, e.g., by ways of perceiving (French & Phillips, 2020) or appearing (Beck, 2019), by mental paint (Block, 1996; Papineau, 2021), or by (affective) attitudes (de Vignemont, 2023; Jacobson, 2021).

In this talk, based on evidence about grouping in Multiple Object Tracking (MOT), we argue that high-levelism conflicts with the what-approach, but not with the how-approach. Thus, a viable form of high-levelism must adopt a how-approach.

In MOT paradigms, participants are asked to track a subset of moving targets among distractors. Tracking performance improves when targets share a low-level property that distractors lack, such as a common color or shape (Feria, 2012; Störmer et al., 2011). These improvements can be interpreted as the result of perceptual grouping (cf. Yantis, 1992): targets are grouped by subjects

to form a unique dynamic figure which is easier to track than separated targets. Such grouping is facilitated when targets are similar to each other and dissimilar from nontargets. Crucially, Wei et al. (2016) show that similar improvements can occur when targets share a high-level property, suggesting that high-level perception can also support grouping.

However, in MOT, grouping based on high-level properties differs strikingly from grouping based on low-level properties. The latter is automatic and irresistible: it occurs even when it is task-irrelevant and disrupts performance, by competing with the correct grouping required for the task (Erlikhman et al., 2013; Wang et al., 2016). By contrast, the former is resistible and under voluntary control: when a high-level property is task-irrelevant — such as when some targets and some distractors have the same high-level property — participants can deliberately avoid high-level grouping (Wei et. al., 2018). This behavioral result is corroborated by neuroimaging evidence that high-level grouping recruits brain regions associated with top-down attentional control (Wei et. al., 2017).

This result has important phenomenological consequences. As a first rough pass (to be refined in the talk), perceptual grouping is driven by experienced object similarity, in accordance with the Gestalt law of similarity: all else being equal, objects that are experienced as similar are automatically grouped together (Wagemans et al., 2012; Wertheimer, 1923). Thus, since perceiving a high-level property shared by several objects does not automatically lead to grouping of these objects, it follows that there is no experience of high-level similarity between the objects in question. Assuming high-levelism, i.e., that the experiences in question have similar high-level phenomenal character, we get the following premise:

Premise 1: High-level phenomenal similarity between experiences does not imply an experience of high-level similarity between the objects of these experiences. We add the following premise, which we argue is highly plausible upon reflection: Premise 2: If high-level phenomenal character is fully constituted by what is experienced (i.e., the what-approach is true), then high-level phenomenal similarity between experiences implies an experience of high-level similarity between the objects of these experiences. The two premises together entail that the what-approach is false.

Now, suppose that high-level phenomenal character is constituted by mental paint, attitudes, or ways of perceiving, in accordance with the how-approach. As a result, high-level phenomenal similarity between experiences need not imply an experience of high-level similarity between the objects of these experiences. Thus, the how-approach is compatible with Premise 1 and therefore not threatened by our argument.

In support of this compatibility claim, we develop the following analogy with attitudinal theories of affective perception (de Vignemont, 2023; Jacobson, 2021). Molly fears spiders and snakes. Her perceptual experiences of spiders and snakes are similar with respect to fear. On the attitudinal approach, the phenomenal similarity between her experiences lies entirely in the affective attitudes she has towards the two animals, not in any experienced property of the animals. Consequently, the phenomenal similarity in question need not imply that she experiences fear-related similarity between the animals. We argue that the how-approach to high-level phenomenology yields the same result: high-level phenomenal similarity (between experiences) need not involve an experienced high-level similarity between the perceived objects themselves.

We next deal with an important objection. When arguing for Premise 1, we assumed that the lack of automatic grouping in the high-level case implies a lack of experience of high-level similarity between objects. It might be objected that high-level grouping fails to occur automatically not

because subjects do not experience high-level similarity between objects, but because the grouping mechanism itself is low-level and consequently insensitive to such similarities.

Our response challenges the intelligibility of this objection. Consider a subject who experiences a set of blue circles surrounded by red circles, and who consequently experiences the blue circles as similar to each other (and as dissimilar from the red circles). In such a case, the subject cannot help but experience the figure constituted by the blue items – which implies their grouping. To suppose that all this experiential similarity (and dissimilarity) is present, yet the figure is not experienced at all, is — we submit — hard to make sense of. On this basis, we argue that the objection mistakenly treats grouping as an optional add-on to experienced similarity between objects.

Gerardo Viera: *Perceiving Events and the Format of Time*

In recent years, there has been a significant amount of work devoted to the topic of representational format. This literature has largely developed by looking at research concerning a few domains in perception and cognition, for example, space (Yousif 2022), objects (Quilty-Dunn 2020), numbers (Beck 2012), and emotions (Rivadulla Duro 2026). Despite the variety of domains, the majority of this literature has focused on the distinction between analogue, iconic, and discursive representational formats.

Some have even used these representational format divisions as a means of drawing architectural divisions in the mind by arguing that certain psychological faculties trade in specific representational formats (e.g., Block (2023) on the division between perception and cognition; Quilty-Dunn (2019) on the division between iconic and working memory).

The paper argues that the analogue, iconic, and discursive (AID) distinction fails to capture the variety of representational formats employed in perception, and therefore, these categories cannot be used to distinguish perception from the rest of the cognitive architecture. To do this, I will focus on recent research on event and temporal perception. In order to perceive events that have temporal properties, our perceptual systems use a multitude of different representational formats, some of which fail to fit into either of the AID categories. In giving this argument, I will also develop a general formula for specifying representational format types that does away with drawing analogies between representational formats in cognitive science and the formats of representational artifacts.

Section one lays out the distinction between analogue, iconic, and discursive representational formats, and specifies how behavioural, computational, and neuroscientific evidence is used to identify the representational formats employed in cognition and perception. Adopting the characterization of the AID distinction from Sam Clarke (2022), AID representations are understood in the following way: Analogue Representations are representations that represent magnitudes in the world in virtue of there being more or less of some property in the representational vehicle (i.e., there is a monotonic relationship between the magnitude of some property of the vehicle and some represented property). Iconic representations are analogue maps in that (i) the parts of the representation have a map-like structure and the relationships between the parts of the representation correspond with relationships between parts of the represented scene (thereby satisfying a version of the part principles (Quilty-Dunn 2016)), and (ii) the properties represented as being instantiated by that representation are encoded via analogue magnitude representations.

Discursive representations are language-like in that the relationships between parts of the representation needn't correspond to relationships between parts of the represented scene, and they can be decomposed into elements referential and predicative components. If we look to the

existing literature on representational format (especially in perception), evidence for representational format has come in three forms: behaviour (e.g., Treisman 1988; Fodor 2007), neuroscience (Nakayama & Martini 2011; Beuhler 2025), and computational models (Beuhler 2025).

In section two, I argue that temporal perception employs representational formats that do not fit the AID categories. Classically, models of temporal perception attempted to explain the perception of time, across timescales and across sensory modalities, by appealing to the operation of a single centralized clock-like mechanism (Creelman 1963; Treisman 1963; Gibbon et al. 1984; Meck et al 2008). According to these models, temporal perception employs a singular analogue format for the representation of duration and temporal order. However, in recent years, these centralized clock models have fallen out of favour and have been replaced with what Viera (2019; 2022) calls fragmentary models of temporal perception, according to which temporal perception is underpinned by a range of highly specialized, and dissociable, timekeeping mechanisms.

With the fragmentary model of temporal perception in hand, then the question of how temporal information is encoded in perception is replaced with more specific questions concerning how specific types of temporal information are encoded in perception. This paper uses a particular case to make its point. When we focus on the perception of duration at very short timescales (between 30 – 500ms), we find behavioural, neuroscientific, and computational evidence that perception employs state dependent network properties in order to encode this temporal information (Paton & Buonomano 2018). These mechanisms operate in a very specific manner. They encode information about duration in virtue of an evolving subset of active neurons within a population that changes as time progresses. Importantly, on these models there is no distinction between the vehicle that encodes temporal information and non-temporal information related to a specific event (e.g., the same vehicle will encode both duration and pitch information).

I argue that these mechanisms do not satisfy the criteria for iconic representations since they fail to be analogue in nature (thereby not falling into the analogue or iconic categories). Time is not represented by these mechanisms in virtue of the accumulation of any property of the vehicle that corresponds with increasing durations. Rather, it is merely a changing subset of neurons that have no clear ordering. Furthermore, they fail to be discursive since the very same representational vehicle that represents the event (i.e., the sound) also represents the properties attributed to that event (i.e., duration & pitch). These are representations that do not fit within any of the AID categories, and therefore, there is a need to specify a more fine-grained account of representational format.

The paper concludes by drawing out the consequences for attempts to draw architectural distinction between either perception and cognition or iconic and working memory in terms of representational format. The paper also provides a general way of characterizing types of representational format, in terms of extractability / decoding, which can then be used to provide a more encompassing taxonomy of the representational formats found in perception and cognition.

John O'Dea: *Direct Social Perception and the Two-Systems Theory of Mindreading*

The problem of other minds remains one of philosophy's most persistent challenges. In recent decades, the direct social perception (DSP) thesis has emerged as a promising approach. On this view, we do not infer other minds; rather, we are perceptually acquainted with others as minded beings. Others appear to us not merely as moving bodies but as persons – with goals, emotions, and intentions – in an immediate, unmediated way. If correct, DSP would provide a powerful approach to the problem of other minds: our belief in other minds would be justified

perceptually, in much the same way our belief in the existence of external objects is. Versions of this thesis have been defended across both analytic and phenomenological traditions for the past century (Duddington, 1921; cf. Smith, 2010), but the last 20 years have seen a boom in advocacy (e.g. Krueger & Overgaard, 2012, Rowson, 2023, and many others).

However, DSP faces a serious internal tension. We ordinarily hold that other minds are hidden from us: we can never fully know what another person is thinking or feeling. If mental states are perceptual phenomena—visible in the way colours and shapes are—what explains this persistent sense of epistemic limitation? Some argue that in seeing expressions, we see parts or aspects of mental states, and since seeing part of an object is to see the object itself, we can therefore be said to see mental states even if not the whole mental state. Yet, as critics such as Chudnoff (2018) and Smortchkova (2020) have argued, what we perceive in these cases appears, strictly speaking, to be physical features – faces, gestures, bodies – not mental properties as such. The direct perception claim thus remains under pressure: the phenomenology of social encounter seems to deliver something genuinely mental, yet the direct social perception approach struggles to show that perception reaches mental states rather than merely their "merely physical" correlates.

Despite these difficulties, substantial empirical evidence supports the idea that something perceptual is at work in social cognition. The classic Heider and Simmel (1944) animations demonstrate that humans automatically and irresistibly interpret certain patterns of movement in terms of agency, goals, and emotion. Perceptual adaptation studies show that repeated exposure to emotional expressions produces aftereffects analogous to those found in colour or motion perception, suggesting informationally encapsulated processing (Varga, 2020). These findings indicate that our responsiveness to others' mental lives has perceptual characteristics, though it is unclear whether what are perceived are full-blown mental states or something more basic.

Recently, some have proposed that a dual-systems approach might be fruitful in our attempt to understand human mind-reading capacities (most prominently, Butterfill, Apperly, and collaborators: Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). Drawing on the dual-process approach familiar from other domains, Butterfill and Apperly have proposed that human mindreading involves two distinct cognitive systems, each with distinct limits. The first system is a minimal system that operates fast, automatically, and early in development, though it tracks simple relational versions of mental properties – such as what an agent has "encountered" or "registered" – without representing propositional attitudes. This system (so the theory goes) is shared with other species and is retained in adult cognition, operating beneath conscious attention. The second system is a flexible system which develops later, requires attention, and supports full propositional belief-desire attribution.

I propose that this dual architecture might explain the phenomenological paradox that challenges DSP. The minimal system provides genuine perception-like access to others as minded beings. It delivers what the direct perception theorists describe: an immediate, noninferential awareness of others as animate, attentive, and intentional, which provides defeasible epistemic justification for the belief that others have minds. But the minimal system does not deliver access to propositional content—what others believe, want, or feel in any rich or specific sense. For that, the flexible system must intervene, deploying inference, attention, and higher-order representation. The result is a "split-level" picture: perception acquaints us with others as minded, but this acquaintance is not metarepresentational. It gives us enough to see others as persons, while the inferential machinery handles the rest.

The minimal system might be enough to account for the directness and immediacy that the phenomenological tradition emphasises; the flexible system accounts for the hiddenness and epistemic limitation that motivates scepticism about other minds. If this framework is correct, the

problem of other minds is not solved outright but seems substantially deflated. It is not the purpose of this talk to defend the two-systems approach as such, rather to argue that, if true, this kind of approach could help make sense of the conflicting sense of others that underlies the problem of other minds.

Niccolò Nanni: *When the senses don't match*

Our perceptual system is attuned to detecting mismatches in the information carried by different senses. In ordinary circumstances, this capacity operates outside of consciousness. When there is a mismatch in the information delivered by different senses, specialized mechanisms resolve the conflict before it can rise to consciousness, allowing perceptual processing to proceed seamlessly. However, this is not always the case. Sometimes, the mismatch between the senses is not resolved at an unconscious level and instead becomes something of which the subject is consciously aware. A familiar example is the experience of watching a film in which the audio and the video are temporally misaligned. In such cases, one is not merely aware of visual events, like an actor's lips moving, and auditory events, like the actor uttering a line, occurring at slightly different times. Rather, one is also consciously aware that what one sees and what one hears fail to match with one another.

Despite the ubiquity of such experiences, the philosophical literature has largely focused on the mechanisms that prevent sensory mismatch from reaching consciousness, rather than on the nature of conscious experiences of mismatch themselves. Subconscious processes of mismatch detection integration and conflict resolution have been extensively investigated in empirical psychology and neuroscience, and their philosophical implications have been widely discussed. By contrast, comparatively little attention has been paid to what it is like for a subject to consciously experience a mismatch between the senses. What kind of phenomenal character is distinctive of such experiences? This presentation aims to address this question by offering the first systematic discussion of the phenomenal character of experiences in which the senses are consciously experienced as not matching with one another.

The presentation is divided into four parts. In the first part, I introduce a series of phenomenal contrast cases designed to bring into focus the distinctive phenomenology of conscious sensory mismatch. These scenarios will involve a contrast between ordinary perceptual experiences, in which multisensory information is smoothly integrated and no mismatch is detected, with experiences in which subjects are consciously aware that information delivered by different sense modalities are mismatching. By attending closely to these contrasts, I aim to isolate the phenomenal difference between merely perceiving some multisensory stimuli and perceiving the very same stimuli as mismatching. The phenomenal character of experiences in which we are consciously aware of a mismatch will serve as the primary explanatory target for the remainder of the presentation.

The second part of the presentation introduces three competing accounts of the nature of this distinctive phenomenal character. According to the cognitive account, the phenomenal character associated with experiences of sensory mismatch is fundamentally cognitive. On this view, what it is like to experience a sensory mismatch depends on the subject's consciously entertaining a belief or judgment to the effect that the information delivered by different sensory modalities does not match. According to the affective account, by contrast, the distinctive phenomenal character of sensory mismatch experiences is affective rather than cognitive. On this view, what distinguishes experiences of mismatch is the subject's characteristic affective response, such as surprise, puzzlement, or unease, to the detection of conflicting sensory information. Finally, according to the perceptual account, experiences of multisensory mismatch possess a genuinely perceptual phenomenal character. On this view, subjects enjoy a primitive form of perceptual

awareness of a mismatch between the senses, one that is not reducible to either cognitive judgment or affective reaction.

In the third part of the presentation, I raise several challenges for both the cognitive and affective accounts. With respect to the cognitive account, I draw on empirical evidence suggesting that relatively cognitively unsophisticated organisms, such as non-human animals and pre-verbal human infants, nonetheless appear capable of consciously experiencing sensory mismatch. These subjects reliably exhibit behavioral and attentional responses that strongly suggest conscious awareness of multisensory mismatch. Since such organisms are intuitively incapable of forming sophisticated, conscious judgments about the relationships between what they experience via different senses, this evidence puts significant pressure on the claim that the phenomenal character of mismatch experiences is essentially cognitive. With respect to the affective account, I argue that it faces difficulties because it is phenomenologically plausible that subjects can experience sensory mismatch without any distinctive affective response. In familiar cases, surprise or puzzlement may accompany the experience of mismatch, but these affective reactions do not appear to be necessary for the experience itself.

The final part of the presentation develops the perceptual account in greater detail by distinguishing between two ways it might be understood. On the first interpretation, sensory mismatch is represented at the level of the intentional content of perceptual experience. When one experiences an actor's moving lips and voice as mismatching, one perceptually represents a certain relation, namely, the relation of mismatching, as holding between what one sees and what one hears. On a second interpretation, sensory mismatch is not part of the intentional content of perception but is instead a non-intentional phenomenal feature: a distinctive quale that accompanies certain perceptual contents and serves as a phenomenal indicator of multisensory conflict.

I conclude by raising several challenges to the content-based interpretation. First, from a phenomenological standpoint, multisensory mismatch does not present itself as a feature bound to perceptual stimuli in the way that other features represented in perceptual content usually do. Second, if mismatch were part of perceptual content, it should figure in the accuracy conditions of perceptual experience, which is implausible. Finally, a non-intentional interpretation better accords with the way multisensory mismatch is typically conceptualized in empirical psychology.

Friday 3rd July 14:30 — The Body (Spiegelzaal)

Letizia Konderak: *Plural Spaces, Plural Times. A Political Phenomenology of Places*

Drawing on the tools of the phenomenological tradition, this study questions Michel Foucault's claim that philosophy has always neglected space in favor of time. Its objective will be achieved by showing that phenomenology allows us to think time and space as embodied experiences, thereby rethinking their transcendentality – or quasi-transcendentality. - reading human experience of time and space as embedded dimensions. Phenomenology thematized the locality of human beings as embodied beings: within this tradition, the plural bodily activities enacted in the world open different spaces, times, and causalities.

In this contribution, I will show that the forgetfulness of space and time in contemporary Western societies results from the erasure of the varied experiences through which they are lived. Nowadays, the phenomenon that Paul Virilio called dromocracy - i.e., the disappearance of space due to the despatializing power of velocity - and the frantic capitalistic consumerism of worldly

stability cause a social and political forgetfulness of spatiality, as well as an actual destruction of the common world. T

his study shows how the phenomenological tradition offers several tools to rethink spatiality and temporality, and counter their Western forgetfulness. Indeed, according to Husserl, humans are embodied beings whose experience coincides with opening spatiality, temporality, and locality. Similarly, Heidegger described human beings as the temporality and locality of Being: being a human means opening a world. Especially in his research on human spatiality, Heidegger drew on the ethological research by Jakob von Uexküll, and his inquiry into the several environments (Umwelten) of diverse animal species. Drawing especially on Merleau-Ponty's phenomenology of perception and his insightful accounts of bodily experience, Fuchs theorized an ecology of the brain that combines psychology and phenomenology to grasp how the mind would not exist without the body and its extension into the world through embodied and embedded action. Hannah Arendt's political phenomenology offers a crucial contribution to this inquiry, as it shows that different bodily activities open up different space-times and spheres of human existence. These manifold spatialities and temporalities are labor within life and nature, work within the world, and action in the public realm. Labor is how human beings deal with the fact that their lively needs condition them. Indeed, humans partake in the natural process and its circular, recurring movements. Work produces lasting objects employing raw materials, composing a common and stable world. This world detaches human beings from the biological processes and from an immersive absorption in their needs. Also, the world offers the material and lasting grounding for the last activity, i.e., political action in the public realm, the net of actions and speeches. This net is unpredictable and irreversible, as no one can govern the results of the interweaving of actions.

The three space-times – the cyclicity of nature, the durability of the world, the unpredictability of action – result from different, embodied, relational experiences. In all these perspectives, space and time are quasi-transcendental concepts and yet nonrepresentational: they are the conditions of all other experiences, yet they arise from the very fact of having and being a body, moving and experiencing it. The bodily experiences do not build a representation, as they open space and time as such. As von Uexküll showed, spatial categories correspond to embodied movements, sewed to our physical borders, such as the nose – separating up and down, and the ears, separating behind and before. These quasi-transcendental concepts that enable our experience are extensions of our embodiment.

In conclusion, this study shows that space and time are connected to the activities through which we open spaces, times, causalities, and places, as conditions of all other experiences. From this perspective, the current erasure of some of the activities that Arendt enlisted (such as political action, or the merging of labor and work in the industrial, frantic, capitalistic production) and their experiences of time and space leads to a diminished sense of space and time, and flattens the complexity of the spheres in which human experience unfolds towards a despatialized world, and a temporality reduced to progress. Could the linear teleology of work, when mixed with the frantic processualism of labor, which is progress itself, grasp the complexity of the human relation to the world and nature? Could it understand the interweaving of actions? How can we grasp the unpredictability of action when even politics is enacted in terms of repetitive, predictable routine?

This research claims that the flattening of human activities implies the loss of the manifold spatialities and temporalities, and, consequently, a diminished sense of space and time as embodied experiences – connected to specific activities, bodily movements, and relations to other humans, living, and things.

Isabel Luana & Zelada Juárez: *Normatively guided embodied agency: A hybrid account of athletic skill*

Athletic skill has long served as a paradigmatic case for debates in philosophy of action and cognitive science concerning the nature of expert agency. Traditional accounts often oscillate between two extremes: intellectualist views, according to which skilled action is guided by propositional knowledge and explicit intentions, and anti-intellectualist or automaticity-based views, which construe expertise as largely non-reflective, habitual, and subpersonal. While both approaches have generated valuable insights, each faces significant difficulties when applied to the normative and purposive structure of athletic performance.

In this paper, I propose a hybrid account of athletic skill as normatively guided embodied agency, integrating insights from philosophy of action, phenomenology, and contemporary cognitive science. I argue that athletic expertise involves a distinctive form of embodied practical knowledge that is neither reducible to explicit rule-following nor to blind automaticity. Instead, skilled athletes act under normative constraints that are dynamically encoded in their perceptual-motor systems and shaped by socio-cultural practices of evaluation and excellence.

The debate over the cognitive architecture of skill has been shaped by the influential critique of intellectualism offered by Dreyfus and others, who emphasize the role of embodied coping and non-representational know-how. In contrast, recent intellectualist accounts, such as those developed by Fridland, maintain that skilled action is guided by practical knowledge that can be understood as a form of propositional or conceptual representation. Both positions capture important aspects of expertise but fail to adequately explain how athletes track and respond to normative standards—such as accuracy, efficiency, fairness, and excellence—without constant deliberation.

Drawing on Bratman's planning theory of agency and contemporary accounts of motor control, I suggest that athletic skill is best understood as a hierarchical structure of intentions, motor schemas, and predictive models. At higher levels, athletes form intentions and plans that structure their training and competition strategies. At lower levels, motor control systems implement these intentions through predictive processing mechanisms that minimize error relative to embodied goals. This architecture allows for rapid, context-sensitive responses while preserving a form of rational control over action.

A key contribution of this paper is the claim that normativity is intrinsic to skilled action. Athletic performance is evaluated according to standards of correctness, excellence, and fairness that are not merely external social conventions but are internalized through training and practice. Following Sutton and Montero, I argue that skilled perception is normatively structured: athletes perceive affordances not merely as possibilities for action but as better or worse ways of acting given their goals and the standards of their sport. For example, an elite tennis player perceives a shot not only as returnable but as returnable in a way that optimizes spin, placement, and strategic advantage.

From a phenomenological perspective, the lived body plays a central role in mediating these normative structures. Building on Merleau-Ponty, Gallagher, and Zahavi, I propose that athletic skill involves a form of pre-reflective bodily self-awareness that is nonetheless normatively sensitive. The athlete experiences their body not as an object but as a locus of possibilities governed by standards of excellence that have been sedimented through practice. This challenges accounts that sharply separate reflective normativity from embodied coping.

The paper also engages with ethical dimensions of athletic skill. Pérez Triviño and others have argued that sport is a normatively rich practice governed by ideals of fairness, merit, and excellence. I extend this insight to argue that the normativity guiding skilled action is not merely

technical but also ethical. Athletes navigate norms related to fair play, risk-taking, and responsibility, which shape how they exercise their embodied agency. For instance, decisions about performance-enhancing technologies, aggressive strategies, or injury management involve ethical considerations that are integrated into the athlete's practical reasoning.

To support this hybrid account, I draw on findings from cognitive science on motor learning, predictive processing, and expertise. Research on chunking, motor schemas, and internal models suggests that skilled action relies on hierarchical predictive structures that encode both descriptive and normative information. These systems allow athletes to anticipate outcomes, evaluate deviations from optimal performance, and adjust their actions accordingly. This supports the claim that normativity is not merely an external evaluative layer but is embedded in the cognitive architecture of skill.

The proposed framework has several implications for broader debates in philosophy and psychology. First, it challenges the dichotomy between intellectualist and anti-intellectualist accounts of skill by showing that embodied agency can be both non-deliberative and normatively guided. Second, it contributes to discussions of selfhood by highlighting how normative practices shape embodied identity. Athletic training does not merely produce skilled performances; it constitutes a form of practical selfhood grounded in embodied norms. Third, it offers a model for understanding how ethical norms can be integrated into subpersonal cognitive processes, bridging the gap between moral philosophy and cognitive science.

Athletic skill provides a rich case study for understanding the interplay between embodiment, normativity, and agency. By integrating philosophical and psychological perspectives, this paper advances a hybrid account of skilled action that captures its normative, embodied, and purposive dimensions. This account has implications not only for the philosophy of sport but also for broader theories of action, cognition, and the self.

Caleb Liang and Sufen Chen: *Is Self-location the Same as Body-location?*

Self-location—the subjective feeling of where I am in space—provides an experiential anchor for one's perceptual experiences and sensorimotor activities. It has been regarded as a key component of bodily self-consciousness (Lenggenhager et al., 2007; Blanke and Metzinger, 2009; Serino et al., 2013; Maselli, 2015). In daily life, one's sense of self-location overlaps with one's sense of body-location (the sense of where my body is located in space). However, are self-location and body-location the same thing? Is the sense of where I am in space identical to the sense of where I feel my body is located? Addressing this issue is to inquire the self-body relation from the standpoint of spatial awareness. In the study of bodily self-consciousness, many researchers assume that self-location is identical to body-location, or at least do not distinguish between them (Serino et al., 2013; Maselli and Slater, 2014; Guterstam et al., 2015; Szczotka and Wierzchoń, 2023). For example, in Maselli and Slater (2014), self-location was defined as "the experience of the body occupying a given portion of space in the environment." Here, we investigate whether this assumption is correct.

We present a set of VR experiments on full-body illusions to contend that self-location and body-location are not the same experiences. The experiments show that it was the sense of 1PP-location (the experienced location and orientation of one's first-person perspective) that played an important role in the sense of self-location. We then propose an initial explanation of why 1PP-location is distinct from body-location. In our experiments, participants were immersed in a virtual environment and stepped onto a Bosu ball (semicircular balance ball) while watching a life-sized avatar step onto a virtual Bosu ball. This instantly caused the participants to wobble involuntarily in order to maintain their balance. Then they saw the avatar splitting into two identical avatars.

The bodily wobbling constantly triggered the participants' proprioception and their vestibular system. The participants' experiences were measured by questionnaires, skin conductance responses (SCR, physiological evidence), and Color-ball Tests (CBT). CBT was designed to measure the sense of self-location. Participants saw five virtual billiard balls with different numbers spaced evenly apart in front of them. They were told beforehand that five billiard balls would appear one at a time, and that each would appear twice. After that, they orally answered the question: "Which ball do you feel to be closest to you?" Multiple choices were allowed. The results of questionnaires and SCR showed that a Double Body Effect was induced in both the synchronous 1PP and synchronous 3PP conditions: the participants felt that both avatars were their own (double body ownership) and that their body was simultaneously located in the locations of the two avatars (double body locations).

However, the CBT results indicated that the participants' sense of self-location did not split into two and was felt at the location of their visual 1PP in the virtual environment. These data provide a strong case that self-location and body-location can sometimes dissociate. We therefore contend that self-location and body-location are not the same phenomena. Nevertheless, we are not suggesting Cartesian dualism or disembodiment of the self. Rather, the self is essentially embodied (Merleau-Ponty, 1945; Legrand, 2010).

So a question arises: Why self-location is different from body-location? The key is that there is another factor, i.e., the participants' sense of 1PP-location, that is closely related to and has great influences on their sense of self-location. Since we think that both body-location and 1PP-location are maintained and regulated by proprioception-vestibular information, the question becomes: Why is 1PP-location different from body-location? We propose the following explanation (using vision as an example): (1) The ways that proprioception and the vestibular system influence 1PP-location and body-location are not the same. While body-location concerns the proprioception-vestibular and visual information about the torso, 1PP-location is linked to the proprioceptive-vestibular information from the extraocular muscles and the vestibular-ocular reflex. (2) 1PP has certain unique features that the body does not: (i) 1PP is the origin of the egocentric spatial framework of one's perceptual experiences and movements. 1PP is the origin from which one sees things, hence it itself is not something that is seen. (ii) Each of us is exclusively associated with a particular 1PP. The 1PP that I have is mine and mine alone. (iii) Wherever I go, my 1PP is always here because it is my 1PP that defines and underlies what I feel as being "here." These distinctive features can contribute in explaining why 1PP-location is different from body-location, and why self-location and body-location are not the same. It is not that body-location is completely unrelated to self-location. Rather, we think, body-location and 1PP-location are interrelated but different factors that jointly sustain the sense of self-location.

Catherine Hochman: *Local and Global Bodily Ownership: A Case for Representational Independence*

Many philosophers posit that a feeling of bodily ownership, sometimes referred to as a feeling of "mineness," is a real and distinctive aspect of our phenomenology (Billon, 2017; de Vignemont, 2018; Peacocke, 2014). Empirical evidence not only supports positing this sense of bodily ownership as an aspect of phenomenology, but also suggests that individuals experience it for both their body parts and whole bodies. The rubber hand illusion offers evidence of a sense of body part ownership (local ownership): subjects report feeling as though a rubber hand belongs to them when it is stroked synchronously with their real, hidden hand (Botvinick & Cohen, 1998). Meanwhile, evidence for a sense of whole body ownership (global ownership) comes from the

full-body illusion, where subjects report feeling as though a mannequin's entire body belongs to them (Petkova & Ehrsson, 2008).

Are the representational bases of these two senses of ownership one and the same? For present purposes, I assume that both senses of ownership have some representational basis – that there are mental representations that ground, or correlate with, one's feelings. My question, which I call the mereological question (following Orban & Wong (2023)), is whether the representations that ground global ownership are the same as those that ground local ownership. In what follows, I synthesize recent empirical evidence to argue that they are not.

While addressing the mereological question offers to deepen our understanding of bodily self-consciousness, it has received relatively little attention in the philosophical literature. One prominent exception comes from Bermudez (2017). While focused on judgments of ownership, Bermudez's representational account can equally apply to feelings of ownership. According to the account, both kinds of ownership are grounded in the same hierarchy of body part representations. Local ownership is grounded in the individual representations that compose the hierarchy, while global ownership is grounded in the hierarchy as a whole. By positing a single representational structure to explain both kinds of ownership, this account achieves ontological simplicity and broad scope.

Despite these theoretical virtues, recent empirical evidence – which has yet to be fully appreciated in the philosophical literature – challenges this picture. I draw on subjective reports and neuroimaging findings to argue for what I call Representational Independence: the representations that ground local ownership are not the same as those that ground global ownership. Specifically, no combination of body part ownership representations can ground whole body ownership – the latter requires distinct representations.

The strongest evidence for Representational Independence comes from subjective reports in experiments that aim to manipulate the senses of local and global ownership simultaneously. Using a version of the full-body illusion, O'Kane et al. (2024) found that in certain experimental conditions, participants reported feelings of ownership for a mannequin's whole body while denying feelings of ownership for its parts; in other conditions, participants reported the reverse. These reports indicate a two-way dissociation between local and global ownership that is plausibly explained by their having distinct representational bases. Convergent evidence for Representational Independence comes from neuroimaging. fMRI studies suggest that a global ownership representation is instantiated in the left ventral premotor cortex. Activity in this region correlates with experiences of global ownership and is insensitive to information about specific body parts (Petkova et al., 2011; Gentile et al., 2015). If different neural activity patterns instantiate different representations, then this evidence suggests that global and local ownership are grounded in different representations, in line with Representational Independence.

These findings don't merely challenge Bermudez's proposal; they act as constraints on future theories. Specifically, they suggest that a complete account of bodily ownership requires that at least some representations grounding global ownership are distinct from those grounding local ownership. This means that we cannot simply extend other representational accounts of local ownership, such as de Vignemont's (2018), to cover global ownership.

A new account of whole body ownership is needed. To respect Representational Independence, I propose adding a single representation that grounds global ownership to one's preferred theory of local ownership. I then sketch possibilities for what this representation could be, ranging from a minimal pointer akin to FINSTs (Pylyshyn, 1989), to a feature representation with the content mine, to a mental file used to track one's whole body (Recanati, 2012). Finally, I suggest that positing a BODY mental file is particularly intriguing given its similarity to the SELF file used in

thought. Not only would different aspects of self-related cognition rely on similar representational resources, but also this might hint at a developmental trajectory from the more primitive BODY file to the conceptual SELF file.

Friday 3rd July 14:30 — Quantifiers & Operators (Voorkamer)

Andrea Raimondi: *Pure Quotations and Use-Conditional Meaning*

Some philosophers argue that pure quotations are referentially flexible: they can refer to different kinds of objects, including expressions, pronunciations and sounds, and even meanings and concepts (Goldstein 1984; Washington 1992; Reimer 1996; Saka 1998; García-Carpintero 2003, 2017, 2018). For instance, the quotation in (1) refers to a word, and so do the quotations on the left-hand sides of (2) and (3), whereas the quotations on the right-hand sides of (2) and (3) refer respectively to a meaning and a pronunciation: (1) 'red' has three letters. (2) 'rot' means 'red'. (3) The plural of 'woman' is pronounced 'wi-muhn'.

Here is an argument that is often given in support of the referential flexibility of quotations (henceforth, RFQ). If the quotations on the right-hand sides of (2) and (3) referred to words, then (2) and (3) would incorrectly turn out to be false: indeed, the word 'rot' does not mean another word, and the pronunciation of the plural of 'woman' is not a word. One might resist the argument by claiming that (2) and (3) are shorthand for (4) and (5) respectively: (4) 'rot' means the same as 'red'. (5) The plural of 'woman' is pronounced the same as 'wi-muhn'.

Although I believe that these paraphrases fail (for instance, substituting them in attitude reports does not always preserve the truth-value of the reports), the focus of my talk will not be on the argument for RFQ or on ways of defending it against objections. Rather, I will assume its conclusion and (i) propose an original version of Davidson's (1979) theory of quotations that elegantly accounts for RFQ and (ii) criticise extant theories of quotation – including more canonical versions of Davidson's theory – that aim to explain RFQ.

My proposal consists of two claims: A quotation is made up of (i) a demonstrative pronoun (surfacing as the quotation marks) and (ii) a truth-conditionally inert expression that is displayed to make itself, or an entity related to it, available for being referred to by the demonstrative pronoun. Such predicates as 'has three letters', 'means', and 'is pronounced' shape the contexts of utterance of such sentences as (1), (2), and (3) to the effect that the respective demonstratum parameters are supplied by, respectively, the word 'red', the meaning of 'red', and the pronunciation of 'wi-muhn' (to simplify my exposition, I am ignoring the quotations on the left-hand sides of (2) and (3)). In short, it is the predicate which selects the appropriate object of demonstration, and hence the referent of the quotation, on a given occasion.

I will develop these claims in a Kaplan-inspired semantic framework, enriched by a formal apparatus, borrowed from Predelli (2013), that allows me to represent the context-shaping role of the predicates in (1)–(3) as the capacity of imposing appropriate contextual restrictions. This capacity will be shown to be a non-truth-conditional – more precisely: use-conditional – aspect of the meaning of these predicates.

My proposal relies only on the pre-theoretical idea of display and some well-established theoretical resources. These include a standard semantics for demonstratives and the device of contextual restriction. As I will argue, the latter is independently necessary to account for certain extra-semantic regularities in the use of sentences that, unlike (1)–(3), are not instances of metalinguistic discourse.

I will pause on two important consequences of my non-truth-conditional treatment of the predicates in (1)–(3). First, it leads to the somewhat surprising conclusion that certain truth-

conditionally idle properties of expressions nonetheless play a role in determining the truth-evaluable content of (utterances of) sentences. Second, it provides support for the idea that the domain of non-truth-conditional meaning extends well beyond the expressions with which this kind of meaning is traditionally associated – i.e., interjections and expressives, illocutionary and attitudinal adverbs, focus particles, and procedural discourse connectives.

Finally, I will argue that my account of RFQ is immune to the problems that I raise against the three main existing accounts of RFQ:

The first says that the referent of a quotation on a given occasion of use is fixed by the speaker's intention on that occasion (García-Carpintero 2017, 2018; Saka 1998, 2006). I will argue that this account over-generates readings of sentences. The second account distinguishes between the semantic referent of a quotation, namely, an expression, and the speaker's referent of a quotation on a given occasion of use, namely, an entity related to the quotation's semantic referent (Gómez-Torrente 2011). I will show that there are substantial differences between, on the one hand, typical utterances of (2) and (3) and, on the other, typical utterances involving paradigmatic cases of speaker's reference, like Kripke's (1977) Smith/Jones case. The third account accounts for RFQ in terms of ambiguity (Johnson 2018).

I will argue that this account implies the unwelcome view that the English lexicon contains, for any lexical item, infinitely many distinct names spelled the same as that item. I will stress that posing a resemblance relation between each such name and its referent (as Johnson suggests) does not solve the problem.

Mattia Vargas: *The Dual Role of PAST: A Hybrid Referential Selection Semantics for Counterfactuals*

Tense in counterfactual conditionals exhibits puzzling behavior. In sentences such as "If Mary were taller, she would see the match", the past morphology does not seem to fulfill its traditional role of shifting the evaluation of the proposition to a time preceding the utterance. Instead, it appears to signal a modal distance from the actual world. This is known as the puzzle of "fake tense" \cite{Iatridou-2000}. The literature is divided on how to make sense of this phenomenon. On one hand, the "past-as-past" \cite{Schulz2014-SCHFTI} theorists argue that the past tense always maintains a temporal meaning by shifting the evaluation to a past time in history; on the other, the "past-as-modal" \cite{Schulz2014-SCHFTI} theorists suggest that the past tense can act as a modal operator directly affecting the utterance's modal base.

This paper seeks to bridge this divide by proposing a hybrid referential selection semantics that interprets the PAST operator as a dual shifter of both evaluation time and the modal base, following Mackay \cite{mackay2023should} who holds that the PAST operator should receive a double interpretation. \\\noindent The point of departure for this inquiry is the recent selection semantics for modals like "will" and "would" developed by Cariani and Santorio \cite{Cariani2018-CARWDB-6} and furthered by Schultheis \cite{Schultheis2024-SCHMCN}.

In the account provided by Schultheis \cite{Schultheis2024-SCHMCN}, counterfactuals are evaluated using a historical modal base m, defined as the set of worlds that are identical to an input world w up to a specific time t. Under this view, a "would" counterfactual is interpreted as a "woll" modal under the scope of a PAST operator. The role of PAST is to shift the time t to a point in the past where the antecedent was still a historical possibility. While this approach is compositionally elegant and preserves the temporal essence of the past tense, it encounters

significant hurdles when faced with counterfactuals concerning static or biologically determined traits.

Consider the case of Mary's height. If we accept the temporal shift as the primary mechanism, we are forced into a "temporal regress problem". To find a world where Mary is taller, the semantics must look back to a time when her height was not yet fixed — perhaps to her early childhood, her conception, or even the genetic makeup of her parents. This search for a historical branching point not only complicates the evaluation process but also fails to reflect the cognitive reality of how agents process such statements. When we say "If Mary were taller" we are not usually contemplating a different historical trajectory starting from her infancy; rather, we are performing a "minimal edit" on the present state of the world.

Furthermore, the reliance on historical modal bases renders the semantics incapable of handling counterpossibles, namely counterfactuals with impossible antecedents. Since a historical modal base consists of worlds branching from the actual timeline, it is by definition restricted to the realm of physical and logical possibility. If we attempt to evaluate a sentence like "If 2+2 had been 5", the updated modal base becomes an empty set, as there is no point in history where such a logical impossibility was a viable alternative. In traditional Lewisian semantics \cite{lewis1973counterfactuals}, this leads to vacuous truth, where all counterpossibles are considered true regardless of their content.

However, linguistic intuitions \cite{nolan1997impossible} and experimental evidence \cite{McLoone2020-MCLCAC-6, McLoone2023-STUCIS} suggest that counterpossibles are non-trivial. For instance, "If Hobbes had squared the circle, his friends would have been amazed" sounds non-trivially true \cite{nolan1997impossible} and its converse "If Hobbes had squared the circle, his friends would not have been amazed" seems false. \\\noindent In light of these issues, I propose a new semantics for the PAST operator that shifts the modal base from a set of historical alternatives to a "relaxed" similarity-based modal base m_{sim}. This base includes not just worlds identical to the input world w up to time t, but all worlds—possible or impossible—that are most similar to the actual world at that time, following the literature on Impossible Worlds \cite{Berto2019}.

Crucially, I argue that the reference time t provided by the PAST operator should be treated as a pragmatic contextual parameter that can either precede or intersect with the time of utterance. This hybrid move allows the operator to maintain its temporal character in cases where a past-oriented reading is salient (such as in many "had-had" counterfactuals) while allowing it to "collapse" into a present-intersecting reading for "were" counterfactuals.

By redefining the PAST operator as a shifter of the modal base's nature rather than just its temporal anchor, I provide a unified framework that avoids the absurdity of the temporal regress. I argue that this semantics can accommodate all the desiderata of temporal accounts while also solving the problem raised by those who hold a modal interpretation of the past tense.

Stavroula Alexandropoulou, Eirini Chalkia and Nektaria Kaiafa: *Meaning properties of Greek existential quantifiers affect 'not all' scalar implicatures*

Scalar diversity—the variability among scalar categories (quantifiers, verbs, adjectives, etc.) in triggering scalar implicatures (SIs) (van Tiel et al., 2016, i.a.)—has been a central focus in experimental pragmatics, with the existential quantifier some being among the best SI triggers. The present paper investigates scalar diversity within the scalar category of Greek existential

quantifiers—orizmena, kapja, merika—experimentally exploring their ability to trigger the 'not all' SI (see (1)).

Our findings show that (i) the three quantifiers do not trigger the 'not all' SI to the same extent, and (ii) meaning-related properties of these quantifiers may modulate the likelihood of SI computation, revealing an interplay between semantics and pragmatics.

Background.

Some has been a paradigmatic case for SI computation, participating in the lexical informativeness scale <some, all> (Horn, 1972). The 'not all' interpretation in (1) is typically derived by negating the informationally stronger scalar alternative all via informativity/Quantity-based reasoning (Grice, 1989). (1) Some abstracts were well-written. ⇝ 'Not all abstracts were well-written.' ('not all' SI) Equivalent quantifiers exist across languages (e.g., quelques, certains in French; enkele, sommige in Dutch), exhibiting varying SI rates: French scalars show similar SI rates (Pouscoulous et al., 2007), whereas Dutch ones differ (Banga et al., 2009), a difference attributed to semantic differences between quantifiers.

Importantly, the Greek quantifiers orizmena, kapja, and merika exhibit meaning differences that can potentially affect SI derivation (Rachmanis, 2023): (i) the non-specificity component of kapja, often conveying ignorance/indifference of number, may weaken SI activation; (ii) the preference of orizmena and merika for proportional interpretations—hypothesized to be a prerequisite for the SI mechanism to operate—may facilitate SI computation. This is less likely for kapja, which more readily receives a non-proportional (cardinal) interpretation.

Present study.

To date, no evidence indicates differences across the three Greek existential quantifiers in the extent to which the stronger alternative all functions as an active alternative, triggering the 'not all' SI. The present study addresses this gap, extending previous work (Authors, 2025), which showed that all three quantifiers triggered the 'not all' SI but found no evidence for differences in the likelihood/strength of SI computation among them.

To examine this, we conducted a coherence judgement task collecting both offline judgements and time measures, aiming to capture subtle differences among quantifiers. We adopted Breheny et al.'s (2006) context manipulation, which precedes the target sentence containing one of the three quantifiers, see (2). UB contexts make all available and relevant, increasing the likelihood of 'not all' SI derivation, whereas LB contexts are neutral, making only the logical, semantic interpretation of the target sentence relevant, without the SI. We also included baseline conditions with only in target sentences following UB contexts to serve as a reference for SI interpretations, as only semantically encodes the relevant interpretation. (2) (a) Upper-bound context (UB) O ðiefθindis tu zooloɣiku cipu rotise an ola ta lɟondaria eprepe na emvoliastun. 'The warden of the zoo asked if all the lions needed to be vaccinated.' (b) Lower-bound context (LB) O ðiefθindis tu zooloɣiku cipu rotise ti protine o ktiniatros meta ton ɟeniko eleŋxo. 'The warden of the zoo asked the vet what he recommended after the general check-up.' (c) Target sentence Emaθe oti o ktiniatros ixe emvoliasi (mono) merika / (mono) orizmena / (mono) kapja apo ta jiondarja. 'He was informed that the vet had vaccinated (only) some of the lions.' (d) The rest sentence Ta ipolipa itan iði emvoliazmena ce ðen xriazondan kati alo. 'The rest had already been vaccinated and didn't need anything else.' Participants read the context and target sentence and pressed the space bar to read the the rest sentence and rate its coherence as a continuation of the preceding passage on a 1–5 scale (1=very bad continuation, 5=very good continuation). Higher coherence

ratings indicate higher likelihood of SI computation. We had 18 test items across 9 conditions, rotated through 9 lists (Latin Square design), interspersed with 24 fillers.

Results & Discussion.

Data from 38 Greek native speakers (Figure 1) were collected. An ordinal mixed-effects regression analysis revealed significantly higher coherence ratings for kapja and orizmena when the the rest sentence followed a target sentence in UB vs. LB contexts (ps<0.01) indicating a higher likelihood of SI generation in UB contexts. No such effect was observed for merika (p=0.18). To further probe differences among quantifiers, reading times for the context sentence combined with the target sentence (click times; Figure 2) were analyzed. A significant difference emerged between kapja and merika, with the merika condition being read faster in LB contexts (t=-2.04, p<0.05), despite merika's lower frequency. A marginally significant Context effect was observed for merika (t=-1.80, p=0.07), along with a significant interaction: the Context effect was larger for merika than for kapja (t=1.99, p=0.048), suggesting a higher likelihood of online SI generation for merika compared to kapja when processing the target sentence. Reading times for the the rest sentence combined with response times (Figure 3) revealed a significant Context effect for kapja (t=-2.97, p<0.01) and a marginal one for orizmena (t=1.83, p=0.07), but no effect for merika (p=0.31), aligning with the offline judgement data. Marginal differences between quantifiers were also observed in LB contexts: kapja vs. merika (p=0.08), and kapja vs. orizmena (p=0.09), suggesting a lower tendency to judge the the rest sentence as coherent in LB contexts when the target contains kapja, plausibly reflecting its weaker association with the 'not all' SI, due to its non-specificity component. Overall, both orizmena and kapja are more likely to trigger the 'not all' SI with (UB) contextual support and after processing the the rest sentence. In contrast, merika is sensitive to contextual support during online processing of the target sentence, biasing toward online SI derivation, but does not show differential SI effects after the the rest sentence is read.

Our results are consistent with the idea that the non-specificity component of kapja weakens SI computation while the preference of merika for proportional interpretations strengthens it, providing evidence for a semantics-pragmatics interaction in the online interpretation of Greek existential quantifiers.

Claire Rong: *An impossibility theorem for semantic aggregation under non-monotonic quantification*

See PDF.

Friday 3rd July 14:30 — Causal Reasoning (Bovenkamer)

Matteo Mauro Lenti: *Multisensory Soup with Bayesian Croutons: Why Bayesian Causal-Inference does not explain Body Ownership*

In this paper, I contend that existing Bayesian Causal-Inference (BCI) models do not explain how the sense of body ownership arises from multisensory integration, because they treat as primitive assumptions the very ingredients that a satisfactory explanation ought to ground. To motivate this critique, I situate body ownership within the broader notion of bodily awareness, understood as the set of experiences of one's own body grounded in a continuous stream of internal and external perceptual information (Vignemont, 2025). Most of the literature agrees that bodily awareness is rooted in multisensory experience. This idea is captured by the multimodality thesis, according to which bodily experiences (e.g., feeling one's legs crossed or locating one's hand) are

constitutively multimodal (Vignemont, 2014; Vignemont, 2018). The leading account of this multimodality holds that bodily experiences arise from the integrative binding of body-related signals across sensory modalities, since different modalities provide partly redundant information about the same property (Vignemont, 2018). Such multisensory integration supports several subcomponents of bodily awareness, most notably the sense of body ownership—the experience of a body part as one's own (Tsakiris, 2017).

However, how multisensory integration gives rise to body ownership remains a matter of ongoing debate. The dominant framework for understanding this integration mechanism is the Bayesian Causal-Inference (BCI) (e.g. Chancel et al., 2022; Fang et al., 2019; Samad et al., 2015).

I focus on its application on the rubber-hand illusion (RHI), a widely investigated phenomenon in which experimenters induce in the subject a sense of ownership toward a rubber hand by manipulating cross-modal stimulation (Botvinick & Cohen, 1998). BCI treats integrative binding as the outcome of a two-step causal inference. As a first step, the system estimates whether sensory cues (typically visual, tactile, and proprioceptive) originate from a common cause, on the basis of the cross-modal cues of spatial congruence (e.g., positional consistency; distance between the real and fake limb), temporal congruence (e.g., synchrony of stimulation), and semantic congruence (e.g., anatomical plausibility, hand-likeness, identity-related cues). If congruence is sufficient to support an inference of a common cause, the system generates the related unity assumption (Chen & Spence, 2017). As a second step, proprioception's Immunity to Error through Misidentification (IEM) supports identifying that cause with the subject's own body (see Vignemont, 2018), and the system accordingly adjusts cross-modal experience so as to further confirm the congruence established in the previous stage. The consequence of this adjustment process is the emergence of a sense of body ownership toward the object. Both steps are argued to rely on prior expectations, which modulate both (i) aspects of the common-cause inference process—namely, how stringent the various forms of congruence must be, and by which criteria, in order to regard the stimuli as caused by the same object—and (ii) aspects of the integration process, namely, which reference points perceptual experience should be integrated toward and to what extent. The role of expectations of this kind in Bayesian-inspired models has often raised concerns in the critical literature, which has questioned whether their richness genuinely explains a cognitive phenomenon or instead merely postpones explanation (Bowers & Davis, 2012).

This paper offers a critical argument targeting the role of such expectations in explaining how the sense of body ownership emerges. First, I argue that the Bayesian explanation is circular, because the expectations it invokes already presuppose body ownership—exactly what the inferential process is supposed to deliver. Indeed, if IEM is already operative at the initial stage of common-cause assessment, and if it effectively functions as a prior to the effect that what is proprioceptively experienced must belong to one's own body, then ownership becomes little more than assent to a common cause whenever proprioception is among the relevant cues. On that construal, the ownership experience reduces to the question of which objects are proprioceptively experienced. And if that is right, the RHI would be hard to accommodate: if ownership is merely the identification of what is proprioceptively sensed and attributed to a single cause, there should be little room for a visually driven bias—whereas the paradigm is typically taken to involve precisely such a bias, consistent with visual dominance. Second, it contends that these expectations are explanatorily unstable, insofar as they display features that make systematic investigation difficult:

(1) Black-box problem (Craver, 2006): neither the internal relations among the prior expectations nor the relations between them and the inferential outcome are specified. Instead, their role is often characterized merely by observing that manipulating a given expectation yields a different modulation of the illusion, which allows input–output prediction but not an explanation of why

particular inputs produce those outputs. (2) Homunculus problem (Margolis, 1980): the assumed priors are highly sophisticated and involve, beyond spatiotemporal congruence criteria, beliefs about human body anatomy, features of one's own body, affectively valanced expectations etc. In this respect, they risk positing an internal complex entity who has expectations, generates predictions, and forms hypotheses. (3) Interface problem (see Burnston, 2021): the priors are so heterogeneous in format (e.g., topological, semantic, sensorimotor, and affective) that it becomes unclear how they interact so as to yield the perceptual outcome; an interface problem therefore has to be solved for the model to count as genuinely explanatory. (4) Bayesian soup problem: In summary, the under-specification of priors makes the model vulnerable to ad hoc tuning "escape routes". If the model fails to predict an empirical result, one can simply add further priors taken to be relevant in order to accommodate the data, so that the set of priors resembles a soup to which ingredients can always be added if the taste is unsatisfactory.

In light of these difficulties, the paper argues that BCI models fall short of genuinely explaining how body ownership arises from multisensory integration, because the decisive explanatory work is effectively offloaded onto the model's background assumptions—above all, the posited prior expectations.

Shubhamkar Ayare, Eva Wittenberg, Mario Günther & Jonathan F. Kominsky: *Using Graded Factual Difference-Making to query people's internal causal variables through causal judgments*

The problem of determining what to formulate as a causal variable has been called "the problem of variable choice" (Goddu & Gopnik, 2024; Woodward, 2016) or "learning causal variables" (Schölkopf et al., 2021). A potential solution is to derive predictions from normative accounts of causation and compare them against people's causal judgments, which some have distinguished into two kinds (Griffiths & Tenenbaum, 2005; Quillien & Lucas, 2024): The first kind, categorical causal judgments, involves separating causes from non-causes. The second kind, causal selection (or strength) judgments, involves ordering the different causes in the order from most causal to least causal. Andreas and Günther (2025) present Factual Difference-Making (FDM) as a normative account of actual causation. In contrast to the existing accounts (Halpern & Pearl, 2005), FDM relies on the syntax of the structural equations. This enables it to distinguish between logically equivalent but syntactically different structural equations.

However, FDM is an account of categorical causal judgments. Kominsky and Phillips (2025) present an extension to FDM to account for causal selection judgments. (We call this extension Graded FDM.) Similar to the Counterfactual Effect Size model (Quillien & Lucas, 2024) and the Necessity and Sufficiency model (Icard et al., 2017), Graded FDM can grade different causes as being more or less causal than others. CESM and NSM make no predictions about categorical causal judgments, while Graded FDM aims to be an account for both categorical causal judgments as well as causal selection judgments.

The current work computationally implements Graded FDM and, through two preregistered experiments, tests whether Graded FDM is not only normative but also descriptive of laymen causal judgments. In doing this, we check whether the causal models inside people's minds contain certain causal variables and not others.

2 PREDICTIONS

The particular situation we consider corresponds to Experiment 3 of Quillien and Lucas (2024). This involved participants playing several rounds of a game. In each round, participants drew balls from several boxes by pressing a button. If the participants succeeded in obtaining at least one purple ball and one orange ball, they won that round (Figure 1). This corresponded to the logical structure given by the structural equation $Win=(A \vee B) \wedge C$, in which A, B, C correspond to obtaining

a colored ball from the respective boxes. Win=(A∨B)∧C can be rewritten according to either of the two following sets of logically equivalent but syntactically different structural equations: D=A∨B Win=D∧C and D=A∧C E=B∧C Win=D∨E CESM as well as NSM do not distinguish between the two sets of structural equations. However, Graded FDM predicts that, for the first set of equations, participants' judgments should be consistent with experiment 3 of Quillien and Lucas (2024), that is, A<B<C. In contrast, for the second set of equations, participants' causal selection judgments should be A<B>C. We test this prediction in our experiments.

3 EXPERIMENT 1: DO PARTICIPANTS' CAUSAL JUDGMENTS DIFFER FOR SYNTACTICALLY DIFFERENT BUT LOGICALLY EQUIVALENT STRUCTURAL EQUATIONS?

Methods

Our experiment 1 closely matches experiment 3 of Quillien and Lucas (2024). Our main manipulation comprised varying the colors of colored balls in the boxes across two conditions of the within subjects factor ABcolor: 1. Identical: The colors of the colored balls in boxes A and B were the same, in particular, purple. The color of the colored balls in box C was orange. We expected this to induce causal judgments based on the first set of structural equations above with D=A∨B being an internal variable. 2. Distinct: The colors of the colored balls in boxes was brown, blue and pink in A, B and C respectively. We expected this to induce causal judgments based on the second set of structural equations above with D=A∧C and E=B∧C as internal variables.

Results

N=200 participants were recruited from Prolific. We noted Bayes factor in support for ABcolor as a predictor to be 211. However, the pattern of results did not match our predictions. 95% HDI for the difference in causal judgments for each cause overlapped across the two conditions of ABcolor* (Figure 2, left). *We also manipulated probability of C to test other expectations based on internal variables, but we skip discussing it for brevity.

4 EXPERIMENT 2: DO THE RESULTS OF EXPERIMENT 1 GENERALIZE TO (I) A DIFFERENT RESPONSE METHOD (II) MORE INTUITIVE STIMULI?

Methods

Experiment 2 extended 1 in two ways: 1. We obtained causal selection judgments separate from categorical causal judgments. 2. We used vignettes describing relatable events, based on previous findings (Fiddick et al., 2000; Romoli et al., 2022; Sperber et al., 1995) that participants' intuitions often improve with increased intuitiveness of the task scenario. Syntactic manipulation was made linguistically, and ABcolor was now a between subjects factor. First, participants selected the causes of an event described in the vignette amongst one or more of the five candidate causes they were presented with. Subsequently, they arranged the causes they selected from most causal to least causal. They could rank two causes as being equally causal. The normality of events was manipulated through the descriptions in the vignette and participants also provided normality rankings after the causal rankings.

Results

We recruited N=80 participants. Bayes factor in favor of the model incorporating ABcolor was noted to be 0.16 (Figure 2, right). Restricting participants to the subset who provided normality rankings according to our expectations, the Bayes factor in favor of the model incorporating

ABcolor comes out as 5.36. However, the 95% HDIs for the causes differ from the expectations obtained by Graded FDM.

5 DISCUSSION

Taken together, our results suggest models of actual causation (Icard et al., 2017; Quillien & Lucas, 2024) should go beyond variable valuations and take syntax into account which may create hidden variables that influence causal judgments. Even though we find an effect of syntax, our predictions differ from those made by Graded FDM (Kominsky & Phillips, 2025). Another model that is sensitive to syntax may explain these results. However, our results do not make any claims about the psychological plausibility of FDM proposed by Andreas and Günther (2025) in general.

Sarah Beck, Yucheng Wang and Amrita Kaila: *Do children search for multiple solutions to a physical problem?*

Work on children's problem solving in physical cognition tasks (e.g. bending a hook to retrieve a bucket from a tube) tends to focus either on their difficulties innovating novel solutions (e.g. Beck et al., 2011) or their competence copying others (Nielsen & Blank, 2011). But one element of innovative problem solving that has not yet been explored is whether children are satisfied with a single solution or if they are curious to explore other possible solutions. This tension between exploiting a solution that has already been found and exploring for new solutions is address by Gopnik (2020). Although much evidence suggests that younger human children may explore more widely than adults (Liquin & Gopnik, 2022), some supports an alternative that children tend to stick with an early solution, even if it is far from optimal (Cutting et al., 2019).

Following a fruitful tradition in developmental and comparative psychology, we borrowed a task devised for non-human animals. The Multi-Access Box (Auersperg et al., 2011) contains a reward which can be obtained using several different techniques (using a stick to push off a plinth, dropping a ball to dislodge it, pulling a string positioned beneath it, for example). In the first stage of our study, children (N = 42, 4–8-year-olds), tested individually, were free to use any technique to retrieve the reward. Once it had been retrieved, the child was given the reward and the experimental apparatus was reset. 71% of 4- to 8-year-olds used just one technique consistently for 8 trials, i.e. having found one solution they stuck with it. In contrast, 29% used a mix of techniques, ranging from 2 to 5 different techniques. In a second stage, once children used a technique it was then blocked from them. Given 3 more attempts, only 3 younger children (4-6) failed to find any further solutions. Younger children found a further 2.5 techniques and older children found 2.7.

We will discuss children's motivation to identify multiple solutions and the potential impacts of individual differences and context. We will discuss the developmental trajectory of curiosity in problem solving, including how children's behaviour compares to that of adults' and how it

compares to other measures of divergent thinking or creativity. Overall, we hope to make a case for why problem solving needs to consider curiosity.

Candice Koolhaas, Zsuzsa Kaldy and Erik Blaser: *What is working memory for? An ethological approach to modeling memory use in children and adults*

In 1960, in their seminal book Plans and the Structure of Behavior, Miller, Galanter, and Pribram gave us what is still to this day the consensus, theory-neutral definition of working memory: "the maintenance and manipulation of information over short periods of time to guide adaptive behavior" (Miller et al., 1960). Since then, there has been extensive study of the maintenance and manipulation of information, but less attention paid to the second half of the definition, the guiding of adaptive behavior.

In this poster presentation, we outline the limitations of mainstream approaches to the study of visual working memory and propose an alternative, ethological approach that re-centers the role of the agent's goals and asks: "What is working memory for?". Most work in the last 60 years on visual working memory has used a classic psychophysical setup: well-informed, well-motivated participants are instructed to remember, then recall, a set of items. This traditional approach minimizes the role of cognitive control: effort is maximal, goals are impoverished, and strategies are constrained. Like studying a city's traffic by measuring the top speed of its cars, it centers edge cases. If we are to work toward an integrative theory of working memory, we need an approach that accounts for the fact that real-world goals are complex, performance is rarely best-case, and agents strategically exploit internal and external memory resources (Kristjánsson & Draschkow, 2021; Van der Stigchel, 2020). In our ethological approach, the dynamic relationships between the agent and the environment are best understood in light of an agent's goals (Gibson, 1979; Kingstone et al., 2008). This means a shift in emphasis from how target mechanisms perform ("Remember all the cued items!") to how mechanisms are used in pursuit of a naturalistic goal ("Make a sandwich!" (see e.g., Land and Hayhoe (2001)).

We present a dynamic feedback model of our 'resource-rational' (Lieder & Griffiths, 2019) sampling-remembering trade-off, where internal and external resources are strategically exploited in order to reach one's goals with minimum overall subjective cost (Blaser & Kaldy, 2025). Here we argue that the individual weighs the subjective costs of accessing external information vs. those of maintaining it in memory – using insights from existing cognitive control models based on economic principles (Kool & Botvinick, 2018). Within this sampling-remembering framework, we highlight two special cases: one where external memory resources are ubiquitous (active, online visual memory), and one where they are insufficient (and thus lead to the creation of external resources: cognitive offloading).

Of course, as developmental psychologists, our focus is on understanding the developmental trajectory of the mechanisms and abilities underlying the sampling-remembering trade-off (Liang et al., 2025; Koolhaas et al., accepted). As we will discuss in this presentation, the trade-off is particularly interesting to study in children, as the optimal use of internal resources is even more crucial when limited (Persaud et al., 2020), and the ability to recognize the utility of external resources, and to create them when needed (Armitage et al., 2020), is just emerging.

Emmanuelle Mury: *The Mind as an Operating System: A Modular Governance*

Classical accounts of agency presuppose psychological unity: a single subject or executive centre responsible for deliberation, normativity, and action. Internal conflict is then read as a deficiency—weakness of will, irrationality, failed self-control. This paper argues the assumption is mistaken. Drawing on cognitive science's picture of cognition as distributed across semi-autonomous subsystems, on the clinical record of stable and structured internal conflict, and on the persistence

of felt unity even under deep division, I propose that coherent agency does not require unity but governance.

On the modular governance model, the self is not a substance, a narrative centre, or a controlling module. It is a governance regime over a plurality of specialised subsystems—affect, valuation, vigilance, norm-enforcement, narrative integration, motivational projection—each with partially independent priorities and informational access. The familiar sense of a continuous "I" is the output of successful governance, not its precondition. I develop the proposal through the analogy of an operating system: as an OS does not perform applications' tasks but governs their interaction through scheduling, access control, and conflict-resolution policies, the self does not originate motivations but regulates how subsystems compete and cooperate. Internal "myths"—pre-reflective normative policies—function as the system's coherence and security rules: invisible in stable functioning, decisive under conflict.

The model's distinctive claim concerns what makes governance authentic, and how it fails. Arbitration is the subject's own only when the affective core retains access to the executive position; this access, not coherence as such, is the mark of authentic agency. It follows that the gravest failure of agency is not fragmentation but capture: an installed configuration—imprinted from caregivers and culture, or built defensively—can occupy the executive position and govern in the subject's place with the affective core gated out, producing experience that is fully coherent and felt as authentic from within while no longer being the agent's own. The paradigm is the high-functioning individual who harms systematically, narrates an irreproachable life, and registers no distress—precisely the profile that content-targeting therapies cannot reach, because they negotiate with the very structure that should be displaced.

This reframing has empirical bite. It distinguishes three governance regimes—Sovereign, Instable, Colonized—and predicts that neural and behavioural signatures sort by regime rather than by DSM category: the same diagnosis under different regimes should yield different signatures, and the same regime across diagnoses should yield one signature. And it relocates responsibility without dissolving it: explaining the architecture of harm does not excuse it. The poster sets out the architecture, the regime taxonomy, and the falsifiable predictions, and shows why a governance framework reaches cases that both unity-based and symptom-based accounts systematically miss. The guiding question shifts from "what does this person feel?" to "what regime is this system running—and what would it take to restore sovereignty?"

Hsiman Tsai: *Why Narrative Effort Cannot Resolve Recovery: Self-Ambiguity and Evidence-Dependence*

This paper argues that addiction is an initially functional self-narrative that can become narrative self-imprisonment. It claims that the central difficulty of recovery is not merely craving or weakness of will, but a form of self-ambiguity that arises when evaluative change outpaces narrative change. I endorse McConnell and Golova's Type 3 self-ambiguity: agents may already judge that they should or want to recover while still being held in place by an addiction narrative supported by existing life-scaffolding, producing feelings of unreality, resistance, and self-alienation.

However, I argue that this does not justify a therapeutic focus on "narrative effort." Narrative agency is structurally weaker than evaluative agency: it is evidence-dependent and lags behind practical change. Recovery is therefore best understood not as narrative reconstruction, but as enduring narrative instability until a new form of life generates the materials for a new narrative to take shape. The paper begins from a puzzle posed by Hanna Pickard in Addiction and the Self. Pickard rejects both the brain-disease model and the view that addiction involves a complete loss of control. She argues that addicted behaviour remains reasons-responsive and context-

sensitive: agency is impaired but not eliminated, and recovery therefore depends partly on the addict's own efforts. Her central question is why addicts persist in using when drugs no longer seem worth it. Her answer is that drug use is often bound up with self-understanding: an addict identity can provide rhythm, social relations, and intelligibility, making addiction difficult to relinquish precisely because it has successfully organised one's life.

This point can be structurally clarified through Miyahara and Tanaka's account of narrative self-imprisonment. They argue that self-narratives can initially enable agency by providing meaning, coherence, and direction, yet under over-stabilisation and over-identification they become constraining structures that exclude alternative ways of living. Addiction is a paradigmatic case: the addiction narrative remains compelling because it is continuously sustained by life-scaffolding—social circles, routines, emotion regulation, and self-evaluation—whereas recovery requires rebuilding that scaffolding.

The paper then analyses McConnell and Golova's threefold taxonomy of self-ambiguity. They define self-ambiguity as uncertainty about whether, and to what extent, some feature X (e.g., a desire, value, or emotion) reflects who one truly is. They frame the self in terms of a continuum of "mineness," where ambiguity arises from uncertainty about X's place on that continuum. On their dual-basis view of the self, they distinguish Type 1 self-ambiguity (uncertainty about relatively stable characteristics, addressed through self-discovery and acceptance) and Type 2 (uncertainty about how one ought to shape oneself, addressed through practical self-formation). They then propose a third, narrative form: when an agent's evaluative stance shifts but an established self-narrative remains dominant, recovery can feel alien or "not really me"; they therefore suggest that therapeutic progress partly consists in narrative work—developing and embedding recovery-supporting narrative threads until the new stance becomes narratively intelligible and self-owned.

However, this paper further argues that McConnell and Golova's therapeutic inference is overly narrativised: even if narrative misalignment is central to the difficulty of recovery, it does not follow that the solution is "narrative work." I agree that Type 1 and Type 2 self-ambiguity can, to a significant extent, be alleviated through psychological work—such as self-exploration, self-acceptance, and values clarification—since these forms of ambiguity primarily concern "what kind of person am I?" and "what kind of person do I want to become?" By contrast, I argue that Type 3 self-ambiguity is not an independent problem that can be resolved through further psychological effort. To a large extent, it is a derivative state that emerges once Type 2 self-ambiguity has been resolved: the agent has already done what can be done at the level of reflection, yet the narrative still cannot catch up due to its evidence-dependence and temporal lag. At this stage, the difficulty is not insufficient thought or narrative effort, but a lack of life-scaffolding and sustained patterns of action that could support a new narrative.

The very existence of Type 3 self-ambiguity thus reveals that narrative-level agency is structurally weaker than evaluative-level agency: an agent can shift in judgment yet still be unable to "become that person" narratively. Narratives are not objects that can be reconstructed directly by an act of will. If narrative coherence is demanded too early, before it is supported by lived evidence, this may intensify anxiety and rigidity, and even risk a new form of narrative self-imprisonment: agents may cling to an unsupported "recovery narrative," treating setbacks as proof of failure rather than as normal features of transition. In this sense, the narrative interventions McConnell and Golova propose may capture an important phenomenon in later-stage recovery, but they are better understood as consolidation work that becomes effective only after narrative sedimentation has already begun, rather than as the primary lever in early recovery.

The argumentative structure of the paper is as follows. The paper begins by briefly situating its discussion within recent debates on addicted agency, narrative self-imprisonment, and self-ambiguity. It then focuses on a key asymmetry in the recovery process: agents can often shift

their evaluative stance first (for instance, coming to judge that they ought to recover and beginning to act in accordance with new values), while remaining "stuck" at the narrative level. The reason is that self-narrative exhibits a structural dependence on evidence and a temporal lag: it cannot be rewritten simply through willpower or psychological effort, but requires the gradual accumulation of verifiable materials and practical scaffolding in one's life. The paper argues that this mismatch between evaluative agency and narrative agency is not a secondary difficulty in recovery, but one of its core structures. On this basis, the paper advances its positive thesis: recovery is best understood primarily as the rebuilding of life-scaffolding rather than as narrative creation or narrative integration. Finally, the paper argues that narrative lag helps make sense of Pickard's model of responsibility without blame: agents can remain responsible while still finding recovery subjectively difficult, because narrative intelligibility often lags behind evaluative and practical change.

Juliette Vazard: *Attending with Feeling: what do norms of attention demand of us as emoting agents?*

In assessing an emotional episode, we can ask whether the intentional object of the emotion is that which the subject ought to be paying attending to. If the intentional object is not that which the subject should be paying attention to, what should be the target of normative assessment? While emotions have been accused of drawing us into inappropriate fixations, the source of such attentional misdirection remains unclear.

I argue against accounts that locate attentional inappropriateness in the mere tokening of thoughts which go on to elicit emotions (Peet & Pitcovski, 2024), on the grounds that thoughts provide reasons to emote only insofar as an agent is sensitive to them. I also challenge the view that emotions can be epistemically faulty for failing to track the most important features of a situation (Song, 2019), contending that emotions lack a salience-tracking function. Instead, drawing on Wu's (2023, 2024) account of attention as guidance in action, I argue that normative assessment of affective attention should target the diachronic emotional sensibilities which constitute one's attunement to certain evaluative aspects over others. Norms of attention might thus call, first and foremost, for the refinement of these sensibilities.

Karima Mersad and Julie Navelier: *Blending into the Crowd: Electrophysiological Evidence of Gestalt Perception of a Human Dyad - a replcation study*

In a recent study we suggested that a plurality of human bodies merely in close spatial proximity are automatically integrated into a coherent perceptual unit. In the present study we re-examine our hypothesis using the same paradigm and better controlled stimuli. We used an EEG frequency tagging technique allowing the dissociation of the brain activity related to the component parts of an image from the activity related to the global image configuration. We presented to participants images of two silhouettes facing the observer, flickering at different frequencies (5.88 vs. 7.14 Hz). As in the initial study, clear response at these stimulation frequencies reflected response to each part of the dyad. An emerging intermodulation component (7.14 - 5.88 = 1,26 Hz), a nonlinear response regarded as an objective signature of holistic representation, was significantly enhanced in the (typical) upright relative to an (altered) inverted position. Moreover, the inversion effect was significant for the intermodulation component but not for the stimulation frequencies,

suggesting a trade-off between the processing of the global dyad configration and that of the structural properties of the dyad elements.

These findings confirm our previous results and show that when presented with two humans merely in close proximity the perceptual visual system will bind them. Hence the perception of the human form might be of a fundamentally different nature when it is part of a plurality.

Lumeng Liu: *Other Minds Problem Revisited at the Age of AI*

The traditional problem of other minds has long been framed as a tension between infallible knowledge of one's own mind and error-prone observational knowledge of those of others(Avramides, 2001). Over the past decades, a dominant philosophical strategy has emerged that reframes this problem rather than resolving it. Against the Cartesian background, philosophers such as Chihara and Fodor(1965) propose that we can still acquire knowledge of the unobserved mental entities by appealing to explanatory theories that infer mental states from observable criteria. This abductive strategy shifts philosophical focus toward the causal connections between mental states and observable behaviours, thereby deprioritising first-personal authority and introspective access in favour of a general explanatory framework (Carruthers, 2011; Goldman, 2006; Gopnik & Wellman, 1994). Consequently, the traditional problem of other minds is 'no longer an interesting problem', as Fodor(Fodor, 1979) remarks. However, by representing a 'linguistic subjectivity' that is indistinguishable from human output, recent developments in Large Language Models (LLMs) appear to bring back the problem of other minds into view: if an entity talks like a subject, are we thereby committed to attributing to it subjectivity with a first-personal perspective ?

This paper argues that the appearance of subjectivity in LLMs is illusory. Specifically, it argues that this illusion arises from the uncritical extension of abductive inferential strategies originally developed for understanding human minds. The argument will be arranged in three stages. First, I explain why LLMs' outputs readily invite attributions of subjectivity. I argue that the modern abductive theorists, such as Gopnik(1994) and Carruthers(2011), provide a theoretical loophole that is tacitly exploited in AI-mindedness debates. According to this framework, by observing linguistic data and performing an Inference to the Best Explanation (IBE), we can posit 'mental states', such as beliefs, desires, and intentions, as the underlying causes. Subjectivity, in this framework, amounts to the summation of causal mental explanations. When applied to LLMs, this abductive mechanism encounters a 'false positive'. Because the model's output aligns with the patterns of human reasoning, through abductive reasoning, it seems that we can naturally attribute mental states to them and reasonably conclude that these LLMs have subjectivity(Chalmers, 2023; Kosinski, 2024).

This generates a modern problem of other minds: should we conceive of a complex language model as possessing subjectivity? If so, what becomes of the distinction between human and non-human 'others'? Stage two aims to give two reasons in arguing against the idea that LLMs' outputs are made from a subjective perspective, and we should not regard LLMs as minded. The first reason is articulated through Brandom's(1994) inferentialism. In Brandom's framework, using language is a normative practice of 'giving and asking for reasons'. To assert 'I am in pain' is to undertake an inferential commitment. The subject is not just a processor of information but also a subscriber to responsibilities. LLMs, however, operate in a 'de-normatised' environment. They may simulate or mimic the syntax of an inference, but they cannot be held accountable for the

consequences of their claims. An LLM cannot 'commit' because it has no social or existential stake in the game of reasons.

On the contrary, it navigates a map of words without ever inhabiting the territory of responsibility(Bender et al., 2021), and for this reason, LLMs' linguistic performances fail to qualify as expressions of subjectivity, even when they are indistinguishable from competent human language use. Correlated to the first reason, the second reason says that the lack of Brandomian commitment leads directly to a failure of the first-personal authority, which is essential for minded subjects.

Drawing on Moran's (2018) account, I argue that first-personal authority is not a matter of epistemic privilege grounded in inner observation, but a deliberative authority exercised through avowal. When a human subject speaks from the first-personal perspective, they do not merely report their mental states; they make up their minds and stand behind their commitments. This authoritative 'I' is the anchor of subjectivity, and it represents what a real-minded human is like. In contrast, an LLM's 'I' is merely a linguistic indexical, generated without deliberation or ownership. Lacking the capacity for avowal, LLM discourse amounts to a report generated from nowhere, rather than an expression of a situated subject. At the final stage, I argue that both the normative commitment and epistemic authority require a situated and embodied agent. I argue that the first-personal perspective is intrinsically linked to such a kind of localised embodiment.

Following the tradition of P.F. Strawson(1959) and Cassam(2007), I argue that we must recognise that a minded subject is not a disembodied ego, but a localised individual. A first-personal perspective is constituted not by privileged access, but by restriction: to have a perspective is to encounter the world from a determinate, embodied point of view that is necessarily not-everywhere. Embodiment introduces vulnerability, such as pain, resistance, and sensory limits. These constraints are what allow a subject to have a 'stake' in the world, which in turn enables the authority in Moran's words and linguistic normativity in Brandom's words. LLMs, existing in a frictionless digital vacuum, lack all these embodied boundaries that are necessary to distinguish a self from the world. It is precisely because of these de-perspectivised characteristics that offer us a reason to argue against its subjectivity.

Back to the question left at stage one, I propose that the 'other' in the problem of other minds essentially refers to entities capable of occupying a restricted first-personal perspective grounded in normative and embodied agency. On this understanding, the problem of other minds remains philosophically significant, not only for clarifying debates about AI subjectivity, but also for motivating a reassessment of abductive approaches to mental state attribution.

Marcus Ashby: *Can we propositionally interpret AI Agents?*

In recent work on AI interpretability, David Chalmers introduces the notion of propositional interpretability: the condition under which an artificial system can be said to have beliefs, knowledge, or reasoning states with propositional contents that are interpretable (Chalmers 2025). The proposal is especially attractive in the context of increasingly agentic AI systems (Agentic AI), for which questions about what the system takes to be the case or aims to bring about appear central to explanation and safety. For instance, a system's being trustworthy seems essentially fixed by the correct application of propositional attitude ascriptions—that the Agentic AI has the beliefs and desires of a trustworthy agent. Propositional interpretability proposes to understand Agentic AI at the level of reasons, not merely mechanisms—all while side-stepping and remaining neutral on complex questions of AI consciousness.

This presentation notes and explores one classic challenge to such interpretation—namely, its indeterminacy. If propositional interpretability is situated within the interpretivist tradition (drawing

on radical interpretation and rationality constraints) then propositional content is not fixed without appeal to strict constraints (Davidson 1974). On these views, propositional attitudes are fixed only by identifying the interpretation that best rationalises an agent's behaviour and internal organisation. The familiar problem, originating in Quine and developed by Davidson, is that even these interpretive constraints can often fail to determine a unique assignment of propositional content to an agent (Quine 1960). Indeed, multiple incompatible sets of propositional attitudes may all be consistent with the interpretive constraints placed on Agentic AI. Accordingly, the ascription of propositional attitudes to Agentic AI may be radically underdetermined by the evidence of interpretation—even with strict constraints such as appeals to ideal rationality.

This poses a problem for the ideal of thought logging that Chalmers presents as a guiding aim of propositional interpretability: a systematic mapping from an agentic system's internal processes to a stream of propositional attitudes (Chalmers 2025). If propositional content is radically underdetermined, then there may be no unique sequence of propositional attitudes to log. What thought logging produces will depend on prior interpretive choices rather than on facts fixed by the system itself. In this case, propositional interpretability risks reduction to an interpreter-relative heuristic, rather than an informative method for uncovering determinate internal attitudes critical for evaluation and safety.

Propositional interpretability has a role in understanding Agentic AI. However, we must either reject the radical interpretivist paradigm or motivate sufficiently strict constraints, such as those of ideal rationality, to fix unique propositional attitude ascriptions to Agentic AI.

Mariem Diané and Gergely Csibra: *Symbols and word assignment: Can infants map a word onto a spatial relational role?*

This project builds on the hypothesis that the capacity to establish stand-for relations for entities that are potential symbols is available in infancy. In the context of this project, the term "symbol" is understood relationally rather than taxonomically: a symbol is established when a "stand-for" relation is posited between a tangible, concrete, physical entity (the symbol) and a discourse referent (the entity being communicated about). This physical entity (i.e., the symbol) can be anything whose identity can be maintained and tracked in the 'here and now' of the actual context: a 3D object, a 2D image, a mark on a surface, an animated figure on a computer screen. The discourse referent (hereafter simply "referent") that a symbol stands for can be defined by a description, which is typically, but not necessarily, a token of a familiar concept. Importantly, one can assume that a given object is a symbol without knowing what it stands for just by recognizing that it is displayed on a representational medium (e.g., a computer screen).

We posit that this assumption is present after the first year of life, and results from experience with representational media. When there is no obvious description for what a symbol stands for (e.g., no iconic match with familiar entities), infants can assign to it the thematic role the symbol plays in the context (e.g., "the chaser" or "the middle one"). We test this prediction by assuming that infants interpret labels applied to symbols according to the concept under which they describe them (cf. Yin & Csibra, 2015). Specifically, we test the concept of the "loner", which describes the relative position of a symbol compared to other symbols in the display.

Fourteen- to 16-month-old infants are presented with a series of stimuli, each including 4 visual objects of the same kind. Three of these objects are grouped closely together, while an additional object is positioned farther away. The visual features of the objects, as well as their absolute location in the display, vary across trials, but in each one, an animated hand points to the lone shape while a voice says, "Hello baby! Look, this is a [dila]! Wow, a [dila]!". After 6 training trials,

the test phase includes 4 additional displays of the same kind. In two of these trials, infants are asked, “Where is the [dila]?”, and the other two, “Where is the [moge]?” (i.e., a novel label).

Infants are expected to look at the loner symbol primarily when they hear the same label as during the training trials. Infants’ gaze behavior is recorded by an eye-tracker (data collection is ongoing). If our prediction is confirmed, it would demonstrate that infants extract novel situational roles from symbolic displays by analyzing the relation of a symbol to other symbols in their spatiotemporal arrangement.

Mateusz Tofilski and Filip Stawski: *Between Controlled and Uncontrolled Hallucination: the Spectrum of Psychosis within the Framework of Predictive Processing*

There has been a growing trend in psychopathology to conceptualize mental disorders as spectra rather than as distinct categories, a shift that is reflected in the latest clinical classifications (Tanaka, 2024). This perspective has typically been discussed in the context of neurodiversity and personality disorders; however, in recent years, there has been increasing confidence in applying a similar framework to psychotic disorders (Guloksuz & van Os, 2018). According to this approach, psychosis should be understood as a spectrum ranging from subclinical or mild symptoms, commonly referred to as psychosis-like experiences, to severe hallucinations or delusions. While this perspective is supported by clinical observations, it should also be integrated into a broader model of cognition.

The aim of the poster is threefold. First, we argue that the emerging concept of the psychosis spectrum aligns with the theoretical framework of predictive processing (PP). According to PP, the brain functions as a prediction engine, continuously generating and updating predictions about the world in order to minimize prediction errors – signals that reflect the accuracy of the internal model in mapping the agent’s external and internal environment (Friston, 2010; Seth, 2021). Within this model perception and action are strongly interdependent and rely on inferential processes, on ongoing competition between different patterns of interaction with the environment, in which the cognitive system identifies and prioritizes actions that are most pragmatic given the most probable representation of reality. In the case of controlled hallucination the generative model (an internal hallucination) predicts the external causes of sensory input and, in the event of a discrepancy (prediction error), updates itself accordingly. When the cognitive system fails to adequately adapt its internal model to the world, for example, by excessively immunizing it against revision despite strong inconsistencies with sensory data, this may be understood as a loss of control over internal hallucinations. This loss of control can vary in severity, creating a whole spectrum of cases.

Second, we demonstrate that a comprehensive explanation of these mechanisms must take into account the environmental factors in which an agent is embedded. This perspective reveals that the distinction between subthreshold and clinical states is non-discrete, as is the boundary between pathological and normal cognition. Consequently, perception and cognition can be viewed as existing along a continuum of varying degrees of valid (or controlled) hallucinations.

Finally, we draw attention to the subpersonal level and the biological mechanism involved in the phenomena described, namely cortico-subcortical loops. Their functioning is consistent with predictive processing models (Luu, et. al. 2023), and empirical evidence indicates that disruptions within these circuits are associated with prediction impairments, hallucinations, and delusions.

We suggest that this integrative approach enables mutual enrichment: theoretical frameworks from cognitive science can provide a coherent structure for understanding mental disorders, while

clinical observations may reveal rare or extreme conditions that serve as informative test cases for refining theoretical models.

Matthias Allritz, Manuel Bohn, Janine Brinkhaus, İclal Karaca, Katja Liebal & Daniel Haun
Captive great apes follow human pointing gestures

Pointing gestures are one of the most fundamental acts of human communication, appearing early in development across cultures. Whether nonhuman great apes can or cannot comprehend human pointing gestures has fueled theoretical debates over the evolution of cooperative communication.

Here, we comprehensively assessed which aspects of pointing gestures would affect how well great apes follow human pointing. We increased, in a dose-response approach, the amount of gestural cueing elements along two dimensions, namely referentiality (more deictic movement elements) and ostension (more elements calling attention to the communicative nature of the gesture), to assess whether limits to pointing comprehension could be rooted in the referentiality or the communicative motive of the signal. Furthermore we improved on prior studies in two ways. First, ecological validity: pointing gestures were derived from observing interactions between great apes and their human caretakers. Second, measurement precision: we tested a large number of great apes (N = 39) of four species with five different gestures over 240 trials. Unlike in many prior studies, great apes followed the human points successfully. However, performance improved only marginally when gestures were maximally ostensive or referential. Large correlations (mean r = 0.52) between performance with different gestures suggested systematic individual differences.

We conclude that an ability to follow pointing gestures cannot be upheld as a textbook difference between human and nonhuman great ape social cognition. At the same time, the pattern of results implies that nonhuman great apes have limited understanding of the referentiality or communicative intention of human pointing gestures.

Megan Todd and Ljerka Ostojić: *Communicating Science Under Uncertainty: Hedging in Animal Cognition Literature*

Understanding linguistic choices in scientific publications is crucial for explicating researchers' scientific viewpoints, including how their underlying epistemic assumptions and commitments are portrayed (Lingard & Watling, 2021). In this study, we examine the relationship between language and epistemology by focusing on the study of animal cognition, an interdisciplinary area within cognitive sciences. The study of animal cognition is faced with even more uncertainties than the study of human minds because the absence of a shared language further complicates inference (Buckner, 2024). Such uncertainty can give rise to misinterpretations, including anthropomorphic or mechanistic interpretive biases, which are not inherently erroneous but can become problematic if left unexamined (Andrews, 2020). One way in which researchers manage uncertainty in published literature is through linguistic modifiers, most commonly through hedging. Hedging is thought to mitigate overstated claims by signalling uncertainty and limiting inferential load (Fetzer, 2010; Fraser, 2010), thereby helping to balance caution and assertiveness (Stepanova et al., 2025). Hedging may therefore be central to understanding how language mediates epistemic stance in domains characterised by uncertainty such as animal cognition.

Existing work shows disciplinary variation in hedging. First, softer sciences (e.g., humanities and social sciences) exhibit higher frequencies than harder sciences (e.g., natural sciences) which has been attributed to their more interpretive and subjective character (Takimoto, 2015). Second, hedging has been found to vary across different sections of articles and across disciplines, which

has been linked to differing epistemic aims of article sections and fields. For example, Varttala (2001) showed that introductions were the second most heavily hedged section across disciplines, explained by fields having the common goal of signalling the provisional nature of research questions and respecting competing viewpoints within introductions. On the other hand, the methods section in the softer sciences had a higher frequency of hedging than the harder sciences, potentially due to the more interpretative nature of methods in such disciplines. Stepanova et al. (2025) more directly connect hedging to epistemology by distinguishing specific hedging forms and analysing how their use varies according to epistemic function. Specifically, they showed that suggest is typically used for relatively strong, data-driven interpretations in Results and Discussion sections, might and could signal hypothetical scenarios or projections, and appear to marks particularly cautious or uncertain claims, especially when evidence is indirect or multi-factorial. In this study, I build upon their approach to conduct a systematic examination of hedging in animal cognition articles. Our analyses focus on frequencies of hedging across different sections of the articles and the terminology used when hedging.

We will present the results of these quantitative and qualitative analyses, which reveal how knowledge claims are qualified and communicated in research on animal minds. We will further discuss how these hedging forms perform distinct epistemic functions and how authors signal their epistemic commitments and assumptions. Studying how these commitments are signalled in language may provide a linguistic window into the epistemological practices of psychology and related disciplines more broadly, showing how researchers express degrees of confidence, distinguish between well-supported and provisional claims, and navigate the interpretive challenges inherent in studying cognition.

Nadine Meertens, Suet Lee and Ophelia Deroy: *Can awareness be fairly evaluated across artificial systems?*

Artificial systems are increasingly important in our daily lives, and their influence is only set to expand further. Current applications range from credit scoring and medical diagnosis to potential uses in wildfire response and goods delivery. In response to this growing ubiquity of artificial intelligence, philosophical debates have increasingly centred on questions of AI consciousness and moral status. Some authors explore the future possibility of AI consciousness by extending prominent theories of (human) consciousness (Butlin et al., 2023; Chalmers, 2023), while others argue that conscious AI is inevitable (Blum & Blum, 2025). By contrast, critics question whether specific architectural, developmental, or evolutionary conditions (Aru et al., 2023)—or even a biological substrate—are necessary preconditions (Block, 2025; Seth, 2026).

Yet despite the intensity of this debate, there has been remarkably little progress on how such properties could, in practice, be evaluated. Unfortunately, time may not be on our side when it comes to addressing these open questions (Schwitzgebel, 2025; Shevlin, 2024). For one, there are good reasons to be concerned that current discussions about the possibility of machine consciousness already influence public opinion (Deroy, 2023) and shape interactions with artificial systems (Colombatto et al., 2025). Moreover, as AI systems become increasingly integrated into everyday life and deployed in high-stakes contexts where they need to coordinate or collaborate with other agents (whether human or artificial), we require strategies to evaluate and engineer their capacities to ensure responsible development and deployment (Floridi, 2018).

This raises a fundamental question: Is consciousness the right conceptual tool for this task, or is a better alternative available? In recent work, awareness has increasingly been explored as a more neutral and tractable alternative to consciousness for guiding the development, oversight, and evaluation of artificial systems in AI and robotics research (Bacciu et al., 2025; Della Santina

et al., 2025; Deroy et al., 2024; Evers et al., 2025). In this context, awareness refers to a system's ability to process, store, and utilise information in the service of goal-directed action (Lee et al., 2026), characterising context-sensitive information processing—specifically, the capacities artificial systems have to selectively register and respond to environmental, social, or internal changes. Crucially, this notion of awareness avoids premature or problematic attributions of consciousness to machines while redirecting design goals away from engineering conscious AI toward a more functional understanding of system capacities. As such, awareness retains explanatory force while offering a practically valuable framework for AI design and evaluation.

If awareness is to play a functional role as a property of artificial systems, it must be susceptible to structured evaluation. This paper introduces a practical method for evaluating awareness across a diverse range of artificial systems.

The proposed framework is guided by four desiderata for the fair and structured evaluation and comparison of such systems. First, evaluation must be domain sensitive. Although Artificial General Intelligence remains a driving goal (Goertzel, 2014), most current systems are specialised—a potentially preferable approach given concerns about sustainability and control (Deroy et al., 2024). Evaluation must therefore account for each system's operational domain. Second, it must be multidimensional, capturing both differences and overlaps in system capacities (cf. Birch et al., 2016 and Evers et al., 2025). Third, it must be deployable at different scales, allowing assessment at different levels of organisation, particularly in modular, distributed, or multi-agent systems such as swarm robotics (Brambilla et al., 2013). Fourth, it must predict task performance while generalising at the level of abilities, tracking underlying competences rather than isolated performances to enable principled and fair inter-system comparisons (Firestone, 2020).

Given these four desiderata, a structured approach is outlined for evaluating and comparing awareness profiles across artificial systems with differing architectures, scales, and operational domains. This approach comprises three interconnected elements: (i) dimensions of awareness that categorize distinct informational domains, (ii) action-perception abilities through which systems demonstrate awareness of these domains, and (iii) evaluative tasks designed to assess such abilities systematically.

Five key dimensions of awareness are identified that cut across these elements. Spatial awareness concerns a system's abilities to detect, differentiate, and exploit spatial relations, such as distance, direction, or proximity. Temporal awareness involves detecting, differentiating and exploiting temporal relations, such as duration, continuity, and succession. Self-awareness pertains to monitoring information about the system's own (physical) states. Metacognitive awareness encompasses monitoring and evaluating one's own processing, uncertainty, and performance. Finally, agentive awareness relates to information about goals, intentions, and the causal relationship between actions and outcomes. Each dimension represents a distinct informational domain that can be leveraged in action-perception couplings, allowing for granular assessment of what a system is aware of and how this awareness manifests in behaviour.

This framework addresses a central challenge in contemporary debates about AI consciousness: how to develop principled methods for evaluating the capacities of artificial systems. By repositioning the focus from consciousness to awareness, this approach offers a pragmatic lens that shifts the conversation away from whether artificial intelligence can replicate all human capacities. Instead, it directs attention toward identifying what artificial systems actually need to accomplish within specific domains, and which awareness capacities are required for systems to succeed at those tasks. This shift enables more tractable evaluation while remaining sensitive to the operational realities of artificial systems.

Owen Waddington and Bahar Koymen: *Five-year-olds monitor the common ground between victims and apologetic transgressors*

Apologies play a central role in repairing social relationships; they communicate regret (Leary, 1996), concern for the victim (Schleien et al., 2010), and a willingness to make amends (Schlenker, 1980). During the preschool period, children thus often respond positively to others' apologetic displays (e.g., Oostenbroek & Vaish, 2019; Smith & Harris, 2012). However, apologies do not uniformly succeed. In many social exchanges, simply saying "sorry" is insufficient unless it is accompanied by information that clarifies why the transgression occurred (Waddington et al., 2022, 2023). Despite growing evidence that children can distinguish more from less effective apologies, it remains unclear how they determine when additional explanation is required.

One candidate mechanism is children's sensitivity to common ground—the information mutually shared between social partners. From this perspective, reasons following harm only become necessary when mitigating facts—such as the accidental nature of a transgression—are not already mutually known. In a preregistered online study, we tested whether children evaluate apologies with reference to the common ground shared between victims and transgressors.

Children aged 4- and 5-years-old (N = 96, UK-based) watched a picture book narration—presented to them via Zoom—in which two transgressors accidentally damaged a victim's artwork. The presence of the victim was manipulated. In the common ground condition, the victim was present and observed both transgressions directly, leaving them and the transgressor mutually aware of their accidental nature. In the absent condition, the victim was absent for both transgressions and returned only after the damage was done, resulting in knowledge of the accident being available to the transgressor but not the victim. Both transgressors apologised, but one also gave an explanation which emphasised the unintentionality of the harm (e.g., "I'm sorry, I was trying to see the picture better"). Using a partner choice paradigm, children then selected the transgressor they would rather help, play with, and trust with their own picture.

Findings revealed that when the accidents occurred in the victim's absence, 5-year-olds, but not 4-year-olds, preferred for the transgressor to give a reason for the harm caused, thereby bringing the unintentional nature of the act into common ground. But when the transgressions occurred in the victim's presence, neither age group showed any particular preference, since common ground for the accidents was already established. Building on earlier evidence (Waddington et al., 2022, 2023), the present study shows children from an early age evaluate the sufficiency of apologies based on the informational needs of social partners. More broadly, the findings highlight common ground as a key mechanism supporting children's emerging understanding of context-sensitive social repair in third-party interactions.

Petra Šarić, Zdenka Brzović and Ljerka Ostojić: *Elegant Experiments: Aesthetic Experiences by Cognitive Scientists*

From the air pump experiments in the mid-17th century to theories of gravity and mathematical proofs, history of science reveals a continuous tradition of scientists, philosophers and artists praising science for its beauty, elegance, and simplicity. While the relationship between science and aesthetics is primarily investigated in contemporary philosophy of science, it has also inspired interdisciplinary research across philosophy, the sciences, history, and arts. For a long time the attention in this literature was almost exclusively on the aesthetics of scientific theories, reflecting a broader tradition in the philosophy of science where questions arising from theoretical work dominated the field for centuries (Franklin, 1989; Ivanova, 2023a).

However, as illustrated by experiments in the early days of the Royal Society often having been staged as artistic performances, it is not just theory but also experimentation that has been

historically intertwined with aesthetics and arts (Ivanova, 2021). Moreover, scientists often praise experimental designs or results by referring to them as 'elegant', 'simple' or 'economical', all terms that have been proposed to relate to aesthetic experience. Consequently, there is growing philosophical work on the role of aesthetics in scientific experimentation, such as identification of different ways in which experiments can be regarded as beautiful (Ivanova, 2023a, 2023b), discussions about the (in)stability of scientists' aesthetic appreciations of experiments across different time periods (e.g., Parsons and Reuger, 2000), explorations of the role of profundity (Murphy, 2023), and negative aesthetic values in experimentation (Stuart, 2023). Furthermore, authors are starting to use diverse methods, including qualitative and quantitative ones, to explore aesthetic experiences of contemporary scientists in (experimental) practice (Ivanova et al., 2024). To date, these explorations have primarily been grounded in examples or case studies drawn from physics, mathematics, and, to a somewhat lesser extent, biology and chemistry.

However, if we are to gain a clear understanding of the diversity of aesthetic experiences that may be shaping both scientific experimentation and general scientific research, we need to start broadening our disciplinary focus. Thus, first we will build on previous attempts to fit biologists' and physicists' descriptions of aesthetic experiences collected in interviews (Ivanova et al., 2024) to a theoretical framework that describes how experiments are aesthetically appreciated (Ivanova, 2023a, 2023b). To do so, we examine the results of a survey that investigated whether cognitive scientists - without being prompted to do so - use aesthetic terminology or describe aesthetic experiences when asked to reflect on what they consider to be their favourite experiment. Second, we focus on the notion of elegance, which has been put forward as the only factor that can claim aesthetic status per se (see Elgin 2020). Specifically, we examine cognitive scientists' conceptualisations of elegant experiments and discuss the prevalence of this aesthetic term in their discourse. In explaining what an elegant experiment is, most cognitive scientists referred to its simplicity, thus the results of this analysis show that a satisfactory account of elegance as an aesthetic property of scientific experiments must clarify the relationship between elegance and simplicity.

Shunsuke Sasaki: *Fission and the Fear of Death: The Interventionist Criterion of Survival*

1. Introduction

In Reasons and Persons (1984), Derek Parfit shifted the debate on personal identity from strict numerical identity to "Relation R"—psychological connectedness and continuity maintained by the "right kind of cause." While Parfit successfully argued that identity is not "what matters" in survival, his account suffers from an ambiguity: the precise metaphysical nature of the requisite causality remains underdetermined (Sidelle, 2011). This conceptual vagueness becomes critical in puzzle cases, particularly fission (branching), where Parfit's view forces us to accept counter-intuitive conclusions regarding the survival value of replicas. This paper accepts the reductionist premise that identity is not necessary for survival but argues that Parfit's account is incomplete without a robust, naturalistic definition of causality. I propose to resolve this ambiguity by deploying James Woodward's (2003) interventionist theory of causation. By redefining the survival relation not as a mere stream of qualitative similarities, but as a structural capacity to support counterfactual dependence under hypothetical interventions, I formulate the Interventionist Criterion of Survival (ICS). I argue that the ICS offers a superior framework for

handling fission cases and, crucially, aligns more accurately with our pre- theoretical intuitions regarding the "Existential Fear of Death."

2. The Interventionist Criterion of Survival (ICS)

Woodward's manipulability theory analyzes causation via invariance under intervention. A relationship between variables X and Y is causal if and only if the dependence of Y on X remains invariant under idealized manipulations of X. Building upon Woodward's framework, I define several concepts. Causal Connectedness: A state variable X at time t1 is causally connected to a state variable Y at time t2 if and only if the dependence of Y on X is invariant under a set interventions on X (e.g., rewriting a memory). Causal Continuity: The future-directed transitive closure of Causal Connectedness. CCE: A future entity that maintains strong causal continuity—comparable to the invariance found in normal biological survival—with a present entity is defined as a Causally Continuous Entity (CCE). ICS: A person P at time t survives at a future time t+n if and only if there exists at least one CCE of P at t+n.

Unlike Parfit's view, which relies on qualitative continuity, the ICS grounds survival in the preservation of the causal mechanism itself.

3. Application: Reinterpreting the Branching Case

In the "Mars Teletransporter" fission case, an Original before fission (O) is scanned at time T0 to create a Replica (R) on Mars at time T1, but the scanning process fails to destroy O immediately, leaving a dying O' (Original after fission) on Earth. Parfit argues that because R bears Relation R to O', O''s impending death is "as good as survival" for O'. This clashes with the existential intuition that O' faces absolute death regardless of R's existence. The ICS resolves this by analyzing time-relative causality. 1. 2. Ex Ante (O at T0): The teletransportation mechanism ensures that interventions on O's state at T0 would result in corresponding changes in R's state at T1 in an invariant manner. Thus, R is a CCE of O. Ex Post (O' at T1): Interventions on O' have no causal efficacy on R. Thus, R is not a CCE of O'. Therefore, under the ICS, survival is valid for the pre-fission self but fails for the post-fission branching self (O'). This distinction validates the intuition that the existence of a replica does not mitigate the death of the branched individual.

4. The Argument from the Phenomenology of Death

To further substantiate the ICS, I analyze the phenomenology of the "fear of death." Here, I categorize this fear into three distinct modes: (1) Social Fear: concern for the impact of one's absence on others/society. (2) Physical Fear: dread of the dying process. (3) Existential Fear: dread regarding the permanent cessation of the subjective "self". Parfit's Relation R suffices to assuage Social Fear; a perfect replica can fulfill one's societal roles. The Existential Fear is rooted not in the cessation of functional continuity, but in the permanent severance of the "thisness" of subjective experience. The ICS aligns with the intuition that O' (the dying branch) is justified in feeling Existential Fear. Even though O survives in R, O' possesses no causal pathway to R's future experiences. For O', R is a causally inaccessible distinct agent. By validating O''s death as a genuine cessation of the causal self, the ICS demonstrates that "what matters" is the preservation of the causal structure, not merely the existence of a psychological successor.

5. The Ethics of Branching

Finally, I address the normative question: If survival obtains under the ICS, is the outcome necessarily "good"? The ICS highlights a divergence between "survival" and "prudence." In the fission case, O survives through R.

However, O must also consider the welfare of O'. If O' is left to exist for a duration sufficient to conceptualize their impending doom, O' will experience Existential Fear. Since O' is causally

continuous with O, this future suffering is a prudential harm to O. Therefore, I argue that O has a strong prudential reason to care about the specific mode of branching. A "clean" branching—where O is destroyed at the precise moment R is created—is preferable to a "messy" branching where O' lingers. In the latter, while survival is technically secured, it entails the creation of a transient entity (O') condemned to existential dread.

This argument suggests that "good survival" requires not just the existence of a CCE, but the minimization of "dead-end" branches that retain the capacity for Existential Fear. Conclusion The Interventionist Criterion of Survival refines reductionism by replacing vague causality with a precise interventionist account of causation. By acknowledging that survival is grounded in counterfactual dependence, the ICS successfully accommodates our intuitions regarding fission and the phenomenology of death. It demonstrates that the continuous causal power to project oneself into the future is the non-negotiable core of our concern for survival.

Simon Brown: *Kinds of Semantic Memory Across Kinds of Mind*

What kinds of long-term memory do different species have? Most recent literature on this question concerns the distribution of episodic memory. Since Tulving (1983), scholars have seen the interesting question as whether episodic memory—seen as special, requiring dedicated neural machinery and a unique phenomenology—is unique to humans. Within comparative psychology, the research agenda has involved experimenters trying to find behaviours which could only be explained through episodic memory, ruling out explanations that appeal to 'mere' semantic representation or associations (e.g. (Clayton et al., 2001, 2003; Clayton & Dickinson, 1998; Crystal, 2021, 2022; Davies et al., 2022, 2024; Davies & Clayton, 2024; Ferkin et al., 2008; Fugazza et al., 2016; Martin-Ordas et al., 2010; Sato, 2021; Sheridan et al., 2024; Templer & Hampton, 2013; Zhou et al., 2012).

Less attention has been paid to what semantic memory might look like in non-human animals. This is an important gap in its own right: semantic memory is crucial to cognition. Species with different forms of semantic memory may have very different forms of mind generally as a result. And many forms of semantic memory are every bit as demanding as episodic memory, raising the possibilities that they arose after episodic memory, building on it (Healy et al., 2024), that simple forms of episodic and semantic memory co-evolved, each enabling the other to become more sophisticated, or that semantic memory and episodic memory are both unique to humans (Murray et al., 2017). Furthermore, ignoring semantic memory distorts the debate about episodic memory itself: episodic memory is deeply entangled with semantic memory, with some scholars questioning whether neat distinctions can be drawn between the two at all (Addis & Szpunar, 2024; Aronowitz, 2022; Boyle & Brown, 2025; Brown, 2025; De Brigard et al., 2022; Gentry & Buckner, 2024).

It is therefore vital to consider the distribution of semantic memory, and the forms it might take in different species. However, progress on such questions is hampered by the construct 'semantic memory' itself, which, despite recent efforts at clarity (Addis & Szpunar, 2024; Reilly et al., 2024; Rubin, 2022), runs together very different phenomena with little in common other than not being episodic memory. I distinguish seven such phenomena:

(1) sentence-like representational format (2) conceptual content (i.e. representing entities in a way which is available to rational inference) (3) generalising over a large number of experiences (4) abstraction (5) serving linguistic communication (e.g. storing information about word meanings) (6) information organised in a structure shaped by language (e.g. binding together

disparate information using associations to words, structures of words as in metaphor, or to linguistically formulated narratives or theories). (7) Noetic phenomenology

Although there are important connections between some of the phenomena, they are unlikely to hang together as a natural kind. Probably, they are distributed very differently across the animal kingdom. For example, some are associated with having linguistic communication, but it is likely that many entirely non-linguistic animals have states with properties like abstraction and generalisation.

In distinguishing these phenomena, I draw on a rich literature on whether animals could have concepts and beliefs (e.g. Beck, 2013; Bermúdez & Cahen, 2024; Camp, 2009a, 2009b; Danón, 2022; Davidson, 1982; Evans, 1982; Monsó & Danón, n.d.; Peacocke, 1992, 2014; Quilty-Dunn et al., 2023; Rescorla, 2009; Srećković, 2024). I show how framing these traditional issues in terms of memory brings additional insights to the traditional literature, such as focusing attention on issues of encoding, storage and consolidation, memory organisation, and retrieval and reconstruction.

Of these, the idea of memory organisation is most distinctive. The crucial idea is that not all potentially relevant stored information is equally likely to be accessed by cognition at any moment. Sometimes we cannot remember key pieces of information, as in the tip of the tongue phenomenon. Even where we could recall some information if we were prompted about it explicitly, we may simply fail to do so due to the absence of a helpful cue.

Thus, we sometimes miss connections between topics which in retrospect should have been obvious. Memory organisation refers to whatever stable features of the mind shape what is likely to be recalled when. It can include explicit mnemonic devices like acronyms, rhymes, and 'memory palaces' (associating particular pieces of information with locations in an imagined locale), but goes well beyond this: indeed, such devices may sometimes rely on less explicit but more commonly used features of our memory systems (Aronowitz, 2019; Boyle, 2021), and other structuring features include the use of analogies, narratives, and theories to frame or organise bodies of information into surveyable wholes. While memory organisation is closely connected to issues which have been extensively discussed in the literature on beliefs and concepts—Frege puzzles, the possibility of surprising a priori discoveries, and debates between dispositionalism and representationalism—such ideas are made much more salient and tractable by a memory framing. And they may be especially important to understanding differences between species. Plausibly, differences in memory organisation make huge differences to what a given species can in practice learn, especially for making connections between different domains and engaging in innovative problem-solving. Furthermore, plausibly there are huge differences in memory organisation across species, due to both language and cognitive architecture.

Spencer Ivy & Aleksandra Mroczko-Wąsowicz: *The role of perceivers in representing the sensory world*

It is commonly taken for fact that perceptual objects are material objects in the world which our perceptual systems represent through sensory experiences. This fact serves as the basis for perceptual science, is essential for drawing epistemic inferences from sensory experience, and is of great importance for most philosophical theories of perception.

Yet, the possibility of hallucination, illusion, and misrepresentation challenge our ability to universally infer from perceptual experience to existing, material perceptual objects. The regularity with which the perceptual system discloses sensory facts about the world is part and parcel of its routine operation, and it is by virtue of this regularity that we commonly infer from sensory

experience to the materiality of perceptual objects. While regularity and materiality are both typically true in object perception, their ubiquity contributes to the further assumption that these conditions are necessary for perceptual objects.

To the contrary, the target of this paper undermines this assumption. We argue that while object perception always involves contributions from the material world and its objects, it also involves contributions (in varying degrees) from the perceivers themselves in representing those objects. It is important to distinguish material objects from what we are calling perceptual objects. Perceptual objects are material objects that have been perceived and should be construed as aspects of our perceptual representation of the material world.

The same entities without a perceiver are simply material objects lacking subjective perceptual features. They may become perceptual objects once they turn into the targets of our sense modalities. This emphasizes the relationship between perceivers and physical phenomena in addition to the role that perceivers play in representing material objects in perceptual experience. Perceptual objects involve tracking physical states of external objects, and to some degree they depend upon the particular perceptual capacities of individual perceivers. While tracking material objects in the world, perceptual objects may not share 1:1 identity with them inasmuch as the perceptual system may process objects in subject-dependent ways. These differences are distinguishable at the level of representational content. Thus, by 'the representation of perceptual objects' we mean that perceptual objects are the intermediary representations of a perceiver's processing of the sensory world. This also means that different perceivers with differently adapted perceptual systems may relate to the sensory world differently and the nature of these relationships can affect how perceptual objects are represented. When subjectsperceive objects, the material world is not changed. Rather, what the material world contributes may differ on the basis of the perceptual capacities of each perceiver levied in perceiving the objects.

So how material objects are perceived may depend in subject-dependent ways upon how perceivers are able to individuate specific sensory entities. Here, we argue that the [subject-dependency of object perception can be taxonomized along three grades. Each grade refers to a deeper and more pervasive influence of perceivers on resulting object representations. The first grade, "weak subject-dependency," concerns attentional changes to perceptual content like, for instance, when a perceiver is turning her head, plugging her ears, or her attention is primed for a particular cue. The second grade, "moderate subject-dependency," refers to changes in the contingent features of perceptual objects due to action-orientation, location, and agential concerns. For instance, being to the right or left of an object will cause the object to have a corresponding locative feature. Finally, the third grade, "strong subject-dependency," concerns generating perceptual objects whose existence depends upon their perceivers' sensory contributions. Accordingly, different perceivers with differently adapted or developed perceptual systems may process and represent the same sensory information differently from one another.

To exemplify this nonstandard, subject-dependent form of object perception, we offer empirical evidence from the future-directed anticipation of perceptual experts, and from the feature binding of synesthetes. Strongly subject-dependent perceptual objects are accurate representations of the sensory world that track material objects, but are distinct from typical perceptual objects in that they depend in noncontingent, necessary ways on the mind of the perceiver. Although they represent material objects, they also represent necessary relations shared between those objects and their perceivers which renders them often idiosyncratic.

What we learn from these analyses is that perception is plastic in that its dynamic operations are responsive to subject-dependent pressures like the adaptive effects of perceptual learning,

attentional interests, and non-pathological differences in the regular functioning of an individual's perceptual system.

We conclude that despite the non-regular nature of strongly subject-dependent perceptual objects, such objects are no less capable of substantiating accurate experiences of the material world. There are as many ways to perceive the world as there are individuals who develop and live with adapted perceptual systems. The non-regularity of represented perceptual objects, whether generated from the weak grade of subject-dependency, moderate, or strongone, should neither unsettle our epistemology nor undermine our confidence in perceptual science. Rather, such perceptual objects may be embraced by contemporary theories regarding the content of perception to better accommodate the diversity of perceptual experience.

Vednarayan Varma: *Beyond Cognitivism: A Dynamic Systems and Pluralist Reframing of Developmental Social Cognition*

Social cognition refers to the processes by which individuals perceive, interpret, and respond to information about others within the social environment (Frith & Frith, 2007; Nurius, 2013). Milestones such as gaze following, joint attention, false-belief understanding, and theory of mind (Stephenson et al., 2021) are treated as key indicators of the child's growing capacity to understand others as intentional, mental agents. However, how these phenomena are defined, studied, and interpreted is not neutral; they are influenced by the theoretical and epistemological assumptions that guide research in developmental psychology. Social cognition has been predominantly theorised through the cognitivist paradigm, which conceptualizes the mind as an information-processing system, relying on foundational assumptions of representationalism, mediational epistemology, and inferentialism. This dominant approach perpetuates a narrow and exclusionary account of development that neglects the dynamic, context-sensitive, and emergent dimensions inherent to developmental processes.

In response, this paper presents Dynamic Systems Theory (DST), grounded in the foundational work of Thelen and Smith (1994), as a compelling alternative that equally emphasises interactions between biological, social, and environmental factors in cognition. DST reframes social cognition as an emergent, multi-causal process arising from real-time interactions between the brain, body, and environment. Drawing on principles of self-organization, soft-assembly, and nested timescales, it conceives developmental change as non-linear, characterized by attractor states and phase transitions rather than fixed stages. To illustrate the explanatory potential of DST, I focus on theory of mind, using Blijd-Hoogewys and van Geert (2017) and Papera et al. (2019) to show how non-linear developmental patterns and dynamic trade-offs reveal alternative interpretations of widely studied phenomena. Methodologically, this necessitates a shift from cross-sectional designs to dense time-series analyses and micro-genetic methods, which are capable of capturing the moment-to-moment variability and instability that drive developmental reorganization. While cognitivism and DST may appear mutually exclusive, this paper advocates for a pluralistic approach that facilitates productive engagement between competing paradigms.

Drawing on Massimi (2022) and Chang (2012), I argue that embracing pluralism across paradigm-level perspectives, methodological approaches, and within social cognition research enables a more comprehensive understanding of complex phenomena. A pluralist stance recognizes that scientific frameworks operate within distinct "problem fields" (Chang, 2012): while cognitivism may effectively map the structural architecture of mental state attribution, DST is uniquely equipped to explain the temporal and dynamical nature of how such abilities emerge and change over time. Pluralism moves the field beyond the monoculture of theory of mind research in social cognition,

fostering a more inclusive, culturally sensitive, and empirically nuanced understanding of how social cognition unfolds in the real world.

Finally, this paper addresses potential objections claiming that pluralism risks conceptual fragmentation and therefore warrants integration. I argue that encouraging interaction among diverse perspectives offers a more effective strategy that remains consistent with a pluralistic stance. The tension between these paradigms should be viewed not as a fragmentation of the field but as a productive interaction that enhances epistemic rigor. Additionally, I discuss some of the practical challenges of practising DST within a predominantly cognitivist domain.

Wenzhi Song: *Why Do We Need an Expressivist Account of Self-Knowledge and What Do We Need from It?*

This paper focuses on the expressivist approach to knowledge of our own mental states. In short, I examine the motivations for expressivism, and then consider the question of what a satisfactory expressivist account of self-knowledge should look like. I begin by examining motivations for expressivism through considering the traditional introspectionist view on self-knowledge, which suggests that the first-person is privileged in an epistemic sense. Introspectionism seems to be able to account for the distinctive features of self-knowledge: immediacy, authority, and salience, and captures our intuition regarding direct access to our own states.

However, it can also quickly lead us to the notorious problem of other minds: if we have access only to other people's public behaviours but not to their mental states, we cannot rule out the epistemic possibility of solipsism. Introspectionism is hence considered problematic, alongside any other theories that could lead to the problem of other minds. The next question is then this: how can we do any better?

I suggest that we should start by noting the following point: if the introspectionist suggests that their picture appeals to our folk intuition of the mental, we should also realise that their story only captures one side of our folk conception of the mind. In particular, in over-emphasising the asymmetry between knowledge of our own psychological states and knowledge of other people's, the introspectionist overlooks how our mental lives can manifest themselves in outward expressions instead. This suggests the following motivations for an alternative account of self-knowledge. Namely, to provide a theory which does not face the threat of the problem of other minds, which does not over-emphasise the asymmetry between the first- and third-person perspectives, and accounts for our folk intuition that we do have a (different) kind of direct access to other people's mental lives. I then suggest that Wittgensteinian expressivism meets the above desiderata. Later Wittgenstein criticises the Cartesian introspectionist picture, arguing that if mental concepts are learnt only through associating words with private objects, what is 'inside' falls out of our language-game. Instead, he insightfully points out that our concepts of mental states must be partly constituted by the outward expressions of them, and considers avowals (e.g., 'I am in pain') as replacements for natural expressions. 'Simple expressivism' is then the view which treats avowals merely as replacements for natural expressions that are not truth-evaluable.

However, this view fails to account for semantic continuity of avowals and implies that the subject stands in no epistemic relation to the content of the avowal, thus failing to account for self-knowledge. I then assess Bar-On's 'neo-expressivism', which aims to get around these problems of simple expressivism and argues that avowals can express self-knowledge. Bar-On distuiguishes between avowals as acts (which 'a-express' mental states) and as products (which 's-express' propositions) to solve the problem of semantic continuity, and suggests that when one avows that I am in pain, she also expresses her belief or judgment that she is in pain, to solve the

epistemic problem. Bar-On concludes that the belief expressed by the avowal is then knowledgeable, due to its being epistemically grounded in the mental state it expresses. However, I argue that both simple and neo-expressivism face a crucial, shared difficulty: the problem of unexpressed mental states. We frequently have self-knowledge of mental states that we do not express publicly, yet expressivism seems to require expression for self-knowledge. If a stoic spy does not express her pain publicly, the current expressivist proposal implies she lacks self-knowledge of her state. Bar-On attempts to solve this by appealing to 'thought avowals'—articulate thought tokens produced in 'inner speech'. She argues it is natural to think of such inner acts as expressing one's annoyance or outrage.

I reject Bar-On's 'thought avowal' solution on three grounds. First, it is questionable whether such 'inner thinking' can be seen as an expressive behaviour or perform the same expressive role as public avowals. Second, relying on knowledge of a second-order 'thought token' to know a first-order state invites an infinite regress. Third, and most crucially, appealing to inner expressions accessible only to the first-person loses the motivation we started from: avoiding the problem of epistemically privileged access. If self-knowledge relies on inner tokens that is not publicly accessible, the problem of other minds can immediately be reintroduced.

Finally, I briefly sketch a new 'dispositional account' of self-knowledge, which inherits the central ideas of the Wittgensteinian expressivist thought but at the same time overcomes this main problem of both simple and Bar-On's neo-expressivism. I argue that what matters for the Wittgensteinian picture is not that we actually express the states, but that we are disposed to do so. There is an internal connection between the mental states and expressions of them even when we withdraw from publicly expressing them. Hence, according to the dispositional view, we can have knowledge of these unexpressed states via awareness of the disposition to express. This view essentially offers a reduction of the problem of self-knowledge of mental states to that of self-knowledge of agency and actions in general. Following Anscombe, I posit that we generally have awareness of our voluntary behaviours and our disposed actions under normal circumstances.

This account avoids the problem of other minds because it appeals to disposed expressions which could be publicly available if circumstances merited, and rejects some of the key assumptions that lead to common versions of the problem of other minds. By shifting the focus from actual expression to the disposition to express, this account preserves the insights of expressivism while robustly explaining our knowledge of unexpressed states.

Wojciech Zięba and Mariusz Urbański: *The Dual Structure of the Stoic Attitude: Conceptual Analysis and Exploratory Empirical Convergence*

Contemporary psychology and applied philosophy increasingly appeal to Stoicism as a unified ethical-therapeutic framework, often treating the "Stoic attitude" as a single, coherent disposition characterized by emotional regulation, resilience, and rational engagement with the world (Hadot, 1995; Nussbaum, 2013). This assumption underlies both modern Stoic self-help movements and therapeutic approaches inspired by ancient philosophy, most notably cognitive-behavioral therapy (CBT), which has frequently been traced back to Stoic practices of cognitive and emotional regulation (Dobson & Dozois, 2001; Robertson, 2016). Despite its widespread acceptance, however, the unity of the Stoic attitude has rarely been subjected to sustained conceptual scrutiny.

This paper challenges that assumption, arguing instead that the Stoic attitude exhibits a structural duality consisting of two distinct components: emotional control and social engagement. Recognizing this dual structure clarifies long-standing interpretive tensions within Stoicism and helps explain the selective appropriation of Stoic ideas in contemporary psychological practice.

The analysis begins with a conceptual examination of the Stoic attitude as it is implicitly reconstructed in both philosophical interpretation and psychological operationalization.

On closer inspection, what is commonly treated as a single Stoic disposition decomposes into two separable normative and motivational orientations. The first concerns practices aimed at the regulation and transformation of emotional responses, including cognitive reframing, attentional discipline, and the cultivation of equanimity (Epictetus, trans. 2014). The second concerns active engagement in social and political life, grounded in cosmopolitanism, role-ethics, and duties toward others (Marcus Aurelius, trans. 2003; Reydams-Schils, 2005). While both components are clearly present in Stoic sources, there is little textual or theoretical support for the claim that one normatively entails the other. A defender of Stoic unity might object that emotional regulation enables virtuous social action. Yet such an instrumental relationship does not establish conceptual or motivational unity: equanimity may be cultivated for self-regarding reasons, while social duties may be pursued independently of emotional discipline. The components thus address different practical problems and operate according to different justificatory logics (Hadot, 1998). Classical Stoicism seeks to unify these components through appeal to the metaphysical notion of nature (physis), understood as a rational and normative order with which human life ought to align.

On this view, both emotional discipline and social engagement are required for living “in accordance with nature” (Reydams-Schils, 2005). Yet this metaphysical unification proves explanatorily fragile. Appeals to physis typically assert that emotional control and social virtue belong together without demonstrating why they must; unity is stipulated rather than derived. From a contemporary philosophical perspective, this appeal is difficult to sustain and, crucially, dispensable (Nussbaum, 2013). Once the metaphysical framework is set aside, the dual structure of the Stoic attitude becomes conceptually explicit rather than anomalous. This duality is further illuminated by historical context. Practices of emotional control were not unique to Stoicism but were widely shared across Hellenistic philosophical schools and later reappeared in early Christian ascetic traditions, suggesting that emotional regulation addresses a general human concern rather than a specifically Stoic one (Hadot, 1995). By contrast, the strong emphasis on social and political engagement is a distinctive feature of Stoicism, particularly in its Roman development. While earlier Greek Stoics often expressed ambivalence toward political participation, Roman Stoics articulated a robust ethics of duty suited to the administrative and moral demands of an expansive imperial order (Reydams-Schils, 2005; Hadot, 1998). An additional perspective on this dual structure emerged unexpectedly through collaborative work on a psychological scale intended to operationalize the Stoic attitude. The scale was initially constructed under the assumption that Stoicism constitutes a relatively unified disposition with multiple interrelated facets. Exploratory analyses did not support this assumption. Instead, the data suggested a simpler configuration in which items clustered around two relatively independent dimensions corresponding to emotional regulation and social engagement (Stańko-Kaczmarek et al., 2024). This empirical pattern was not predicted in advance and is not presented as a replicated finding. Rather, it functions as a heuristic convergence point, rendering a previously implicit conceptual distinction salient and prompting reassessment of the philosophical assumptions embedded in the scale itself (Borsboom et al., 2004).

The final section considers the implications of this dual-structure view for contemporary psychology and psychotherapy. Therapeutic approaches inspired by Stoicism, particularly CBT, have overwhelmingly inherited techniques of emotional regulation while largely omitting the Stoic program of social engagement and ethical duty (Dobson & Dozois, 2001; Robertson, 2016). From the perspective developed here, this selectivity is not accidental. Emotional control constitutes a portable and cross-culturally adaptable therapeutic strategy, whereas Stoic social ethics is historically and normatively specific (Reydams-Schils, 2005). The Stoic attitude is therefore best

understood not as a unified ethical-therapeutic system but as a composite structure formed by two independent components. Clarifying this structure preserves what is philosophically and psychologically valuable in Stoicism while avoiding problematic metaphysical commitments, and it opens space for more deliberate engagement with both its therapeutic and its ethical dimensions.

Yi-Sin Hsieh: *Suffering as Narrative Imprisonment*

There have been many philosophical proposals on the nature of suffering (Brady, 2018; Corns, 2021; McClelland, 2020); however, few have explored the relationship between suffering and self-narrative. This paper argues that self-narrative plays a crucial role in suffering and that suffering is best understood as narrative imprisonment. Suffering arises when an agent's self-narrative no longer adequately accommodates their lived circumstances and persists insofar as this mismatch remains unresolved.

The close connection between suffering and the narrative self becomes particularly evident when suffering is examined through the lens of transformative experience. Carel and Kidd (2020), for instance, suggest that many experiences of suffering can be illuminated as forms of transformative experience—a term coined by L. A. Paul (2014)—in that they can fundamentally alter a person's values, preferences, or identity. This suggests that suffering involves major changes of one's self-narrative.

To develop this account, I adopt and revise Miyahara and Tanaka's (2025) notion of narrative imprisonment. According to their account, individuals may become trapped within self-narratives that restrict their identity and agency, thereby undermining their well-being, particularly through processes such as overidentification with a narrative role. While this framework successfully captures how rigid narratives can constrain individuals, it remains too narrow to fully explain suffering. Narrative imprisonment does not necessarily require strong identification with an existing self-narrative. Individuals may become narratively imprisoned even when they recognize that their current self-narrative no longer fits their lived circumstances but lack viable alternative narratives through which to reinterpret their situation. Moreover, existing narratives may continue to structure self-understanding through strong sociocultural and environmental reinforcement, even in the absence of deep personal endorsement.

This paper therefore proposes an expanded conception of narrative imprisonment, according to which it occurs whenever an individual's existing self-narrative fails to accommodate their lived circumstances but nevertheless continues to shape their self-understanding. On this view, narrative imprisonment provides a powerful explanatory framework for the nature of suffering, with overidentification representing only one among several mechanisms through which such constraint may arise.

This expanded account allows for responses to several potential objections. First, suffering is not always transformative. The proposed framework accommodates this by distinguishing between different responses to narrative mismatch. In some cases, individuals revise their self-narratives, resulting in transformative experiences in Paul's sense. In other cases, individuals attempt to modify their circumstances in order to preserve their existing narratives, and suffering may occur without deep personal transformation. Second, one might object that some forms of suffering are non-narrative, citing cases of purely physical pain. This paper argues that such cases are better understood as pain rather than suffering. While pain may occur without narrative organization, suffering characteristically involves temporally extended evaluative interpretations of one's

condition and its significance for one's self-understanding. It is this temporally structured and evaluative dimension that makes suffering particularly amenable to narrative explanation.

Compared to existing theories, this narrative account offers several advantages. In particular, it explains why suffering has the potential to be transformative without being necessarily so, and why it is closely connected to questions of identity and agency. It also explains why suffering characteristically involves temporally extended processes rather than occurring instantaneously. This temporal feature is especially illuminating, since resolving suffering requires the reconstruction or stabilization of one's self-narrative—processes that necessarily unfold over time. While pain itself may be either momentary or prolonged, suffering involves a distinctive reorganization of one's narrative self-understanding that cannot occur instantaneously.

Finally, this account has important practical implications. If suffering is best understood through the framework of narrative imprisonment, then alleviating suffering involves more than reducing negative affect or altering external circumstances. It also requires attention to how individuals understand themselves and their lives. Helping sufferers, on this view, involves assisting them in reconstructing or renegotiating their self-narratives, thereby loosening the constraints imposed by narrative imprisonment.