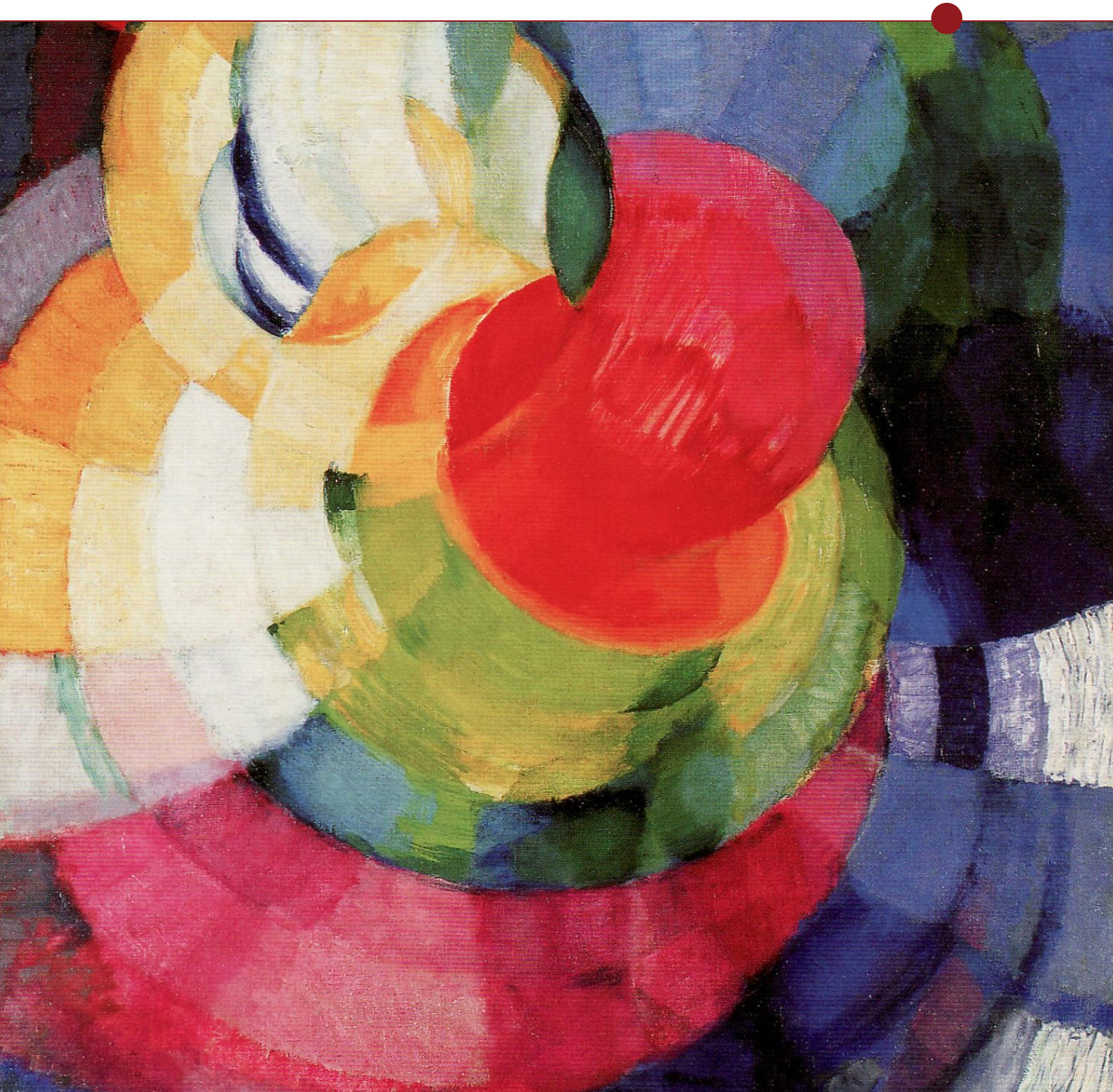
UNIVERSITY OF COPENHAGEN
FACULTY OF HUMANITIES

HOW WE THINK

Rationality and Agency in Thought

FREDERIK T. JUNKER



How We Think

Rationality and Agency in Thought

FREDERIK T. JUNKER

A dissertation submitted to the PhD School at the Faculty of Humanities,
University of Copenhagen in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Philosophy

Thesis supervisor:
Prof. Thor Grünbaum

Submitted on August 31, 2025
Word count (excl. references): 48,491

UNIVERSITY OF
COPENHAGEN

Funded by an open 3-year PhD scholarship
awarded by the Department of Communication,
University of Copenhagen.

Cover illustration:
‘Disks of Newton (Study for Fugue in Two Colours)’ (1912)
by František Kupka.
In the collection of the Philadelphia Museum of Art.
The image is in the public domain.

University of Copenhagen
Department of Communication
Section for Philosophy

Contents

Abstract

Combining philosophical analysis with findings from cognitive science, this dissertation seeks to advance our understanding of the nature of thought. It is argued that new insights can be gained from examining forms of thought often viewed as peripheral to rational and active thought. A closer inspection of such phenomena reveals that traditional distinctions should give way to a more pluralistic view of rationality and agency in thought.

The first article investigates the role of mind-wandering in planning. While mind-wandering can facilitate exploration of alternative goals and actions, excessive exploration risks destabilizing the intentions that guide long-term planning. The article proposes a model of the planning-related functions of mind-wandering that reconciles its exploratory benefits with the need to maintain stable intentions in the pursuit of long-term goals.

The second article examines the agentive status of mind-wandering. Though often described as passive and unguided, mind-wandering also advances our goals. This contribution is explained by mechanisms that monitor, evaluate, and regulate mental representations and processes. Mind-wandering is thus an actively guided learning process. This has important implications for our understanding of rational inference and mental agency.

The third article explores the role of cognitive maps in reasoning. It is shown that cognitive maps can mediate mental transitions that yield conclusions about what is the case or what to do. These transitions qualify as reasoning: conclusions are responses to premise-states; transitions are responsive to rational norms; and the reasoner takes their conclusion to be supported by preceding states and operations. Because these transitions are not rule-governed operations over propositional attitudes, they depart from traditional accounts of reasoning.

The fourth article identifies two versions of the predictive processing framework: one reduces beliefs and desires to a single construct, while the other introduces a distinction between these states. The latter aligns with standard accounts of action and motivation in philosophy and cognitive science, while the former entails a substantial revision. This reveals a dilemma: parsimony of constructs versus a complete explanation of agency and the mind that recognizes the distinct roles of beliefs and desires.

As a whole, the dissertation sheds light on how we conduct our mental lives. By examining the mechanisms that give rise to diverse forms of thought, it deepens our understanding of why our minds are prone to wander, how we plan and reason, and of the forms of control we exert over our thinking. This also yields new insights into the general architecture of the mind.

Resumé

Ved at kombinere filosofisk analyse med indsigter fra kognitionsvidenskaben tilsigter denne afhandling at fremme vores forståelse af tænkning. Afhandlingen viser, at nye indsigter opstår, når vi undersøger former for tænkning, der ofte betragtes som perifere i forhold til rationel og aktiv tænkning. En nærmere undersøgelse af sådanne fænomener afslører, at traditionelle skildringer bør vige for et mere pluralistisk syn på rationalitet og handling i tænkning.

Den første artikel undersøger tankevandringens rolle i planlægning. Mens tankevandringen kan udforske alternative mål og handlinger, risikerer overdreven udforskning at destabilisere de intentioner, der styrer langsigtet planlægning. Artiklen fremfører en model for de planlægningsrelaterede funktioner ved tankevandring, som forener dens eksplorative fordele med behovet for at opretholde stabile intentioner i forfølgelsen af langsigtede mål.

Den anden artikel behandler tankevandringens handlingsmæssige status. Selvom tankevandring ofte beskrives som passiv og uden styring, bidrager den også til at fremme vores mål. Dette bidrag forklares gennem mekanismer, der overvåger, evaluerer og regulerer mentale repræsentationer og processer. Tankevandring er således en aktivt styret læringsproces. Dette har vigtige implikationer for vores forståelse af rationel slutning og mental handling.

Den tredje artikel undersøger kognitive korts rolle i ræsonnering. Det vises, at kognitive kort kan mediere mentale transitioner, der fører til konklusioner om, hvad der er tilfældet, eller hvad man bør gøre. Disse transitioner kvalificerer som ræsonnering: konklusioner er svar på præmis-tilstande; transitionerne overholder rationelle normer; og den ræsonnerende person opfatter sin konklusion som understøttet af de forudgående tilstande og operationer. Fordi disse transitioner ikke er regelstyrede operationer over propositionelle attituder, adskiller de sig fra traditionelle forståelser af ræsonnering.

Den fjerde artikel identificerer to versioner af predictive processing-frameworket: én reducerer overbevisninger og ønsker til et enkelt konstrukt, mens en anden indfører en sondring mellem disse tilstande. Sidstnævnte stemmer overens med standardforklaringer af handling og motivation i filosofi og kognitionsvidenskab, mens førstnævnte indebærer en væsentlig revision. Dette afslører et dilemma: simple konstrukter versus en fuldstændig forklaring af handling og sindet, som anerkender de særskilte roller, som overbevisninger og ønsker spiller.

Samlet set belyser afhandlingen, hvordan vi styrer vores mentale liv. Ved at undersøge mekanismerne bag forskellige former for tænkning, udvider den vores forståelse af, hvorfor vores tanker ofte vandrer, hvordan vi planlægger og ræsonnerer, og hvilke former for kontrol vi udøver over vores tænkning. Dette afføder også nye indsigter i sindets generelle arkitektur.

Acknowledgments

This dissertation would not have been possible without the support, guidance, and encouragement of many individuals and institutions. I am especially grateful to my supervisor, Thor Grünbaum, whose mentorship has been invaluable in shaping both this dissertation and my development as a researcher. His insight and good humor have made our many conversations both inspiring and enjoyable. I have long admired the generosity he shows toward his students. Thor has consistently provided thoughtful guidance on long and often unwieldy manuscripts, and his efforts to foster a collaborative and engaging academic environment have made being part of his orbit a truly enriching experience. I could not have wished for a better supervisor. I am also grateful to Prof. Klemens Kappel, Prof. Myrto Mylopoulos, and Prof. Nicholas Shea for agreeing to serve on my thesis committee and for their willingness to engage with this work.

The project has benefited immensely from the academic community at the University of Copenhagen. I am indebted to my colleagues for the many conversations, reading groups, research group meetings, and informal exchanges. These have made what could otherwise have been a fairly solitary endeavor feel like a shared journey. I am especially thankful to the members of the Cognition, Intention and Action (CoInAct) research group, whose insights and perspectives have enriched me and my work in countless ways. It has been a joy to witness how much philosophers, psychologists, and neuroscientists can gain from working together. Thank you for providing me with an intellectual home—I feel truly fortunate to have worked in a place where such interdisciplinary engagement is not only possible, but encouraged.

I would like to express my sincere gratitude to my peers at the University of Copenhagen, Andrea, Andreas, Anne-Sophie, Bodil, Ditlev, Frederik, Jelena, Kasper, Laura, Maria, and Martin. Their friendship and encouragement helped me navigate the challenges and celebrate the successes of doctoral life. I have cherished the many inspiring discussions and joyful moments we have shared. The camaraderie made the long days at the office not only more manageable, but also far more enjoyable. Thank you for creating a supportive and cheerful environment. I would also like to thank my co-author and co-teacher, Jelle Bruineberg, for many helpful and enjoyable conversations.

Special thanks go to Maria Fülberth, Malthe Hendrickx, Victor Lange, Yair Levy, Andrew Mo, Søren Overgaard, Damiano Ranzenigo, Tessa Super, and Iwan Williams for reading earlier drafts of the articles contained in this dissertation. The manuscripts have benefited immensely from their generous feedback. As have I from their company.

I am also grateful to Felipe De Brigard and the members of the Imagination and Modal Cognition (IMC) Lab and Department of Philosophy at Duke University, where I spent a semester as a visiting researcher. I have many fond memories from my time at Duke and feel very fortunate to have gotten to know and learn from so many wonderful people during my time there. Thanks to Deborah, Gabriela, Kevin, Kaylee, Laura, Lindsay, Mary, Nina, Ricardo, Tzvetan, Victor, and Vijay for making my time there so special. I am grateful to Felipe for including me as part of the team, for his inspiring mentorship, and for introducing me to the wonders of the hippocampus (and the rest of the brain). Thanks to Mig and Mike for letting me into their home and for being such warm and welcoming hosts.

Finally, I would like to thank my friends and family for their unwavering support and encouragement over the years, no matter how inscrutable my job may sometimes seem. Many thanks to my two cats, Nala and Bagheera, whose writing contributions (walking across the keyboard), reminders to take breaks, and occasional mischief offered a much-needed supply of distraction, comfort, and routine throughout the long hours of writing. To my wife, Clara, thank you for your love, patience, and support throughout this journey. I cannot imagine these past years without you. I sometimes joke that, according to you, I was an expert on mind-wandering before I read the first paper on the topic. While I may tend to get lost in thought, know that nothing fills those thoughts more than my love for you.

Introduction

What does it mean to think? This may sound like a suitably abstract question for a philosopher to confront—perhaps even a slightly self-indulgent one. This sentiment is not without merit. Yet thought is also more than a mere philosophical curiosity: it is how we make sense of and navigate the world. Thinking enables us to draw conclusions about how the world is and how we should act within it. When done well, we call it rational. Thinking is also frequently something we do, rather than something that merely happens to us. But what does it mean to think rationally? And in what sense is thinking an action? Although we engage in it constantly, often without noticing, thinking is a deceptively complex phenomenon. Understanding the nature of thinking requires uncovering how various mental structures combine to form the thoughts that populate our minds and enable us to draw conclusions. It arises from the interplay of many bits of intricate cognitive and neural machinery. Philosophy has a venerable tradition of identifying general features of thought, while the sciences of the mind and brain reveal the mechanisms that sustain it. To make progress, we need the perspectives of both.

Traditional philosophical accounts have often focused on paradigmatic cases of rational and active thought such as intentional deliberation and logical inference over propositional attitudes. Yet much of our mental life does not fit neatly into these categories. We routinely engage in forms of thinking that are less reflective, non-intentional, mediated by non-propositional representations, and not governed by explicit logical rules. Far from being a mere sideshow, these modes of thought play a central role in how we make sense of and navigate the world. Investigating them challenges received views about rationality and agency, and reveals that these take more diverse forms than is often recognized. Sometimes we also discover that capacities recognized by traditional theories can be supported by unforeseen mechanisms. By attending to the full range of mental phenomena—including those that seemingly resist traditional classifications of rational and active—we gain a richer understanding of what it means to think and new insight into the cognitive architecture that underpins our mental lives.

This dissertation seeks to deepen our understanding of the nature of thought by attending to cases at the margins of what is traditionally considered rational and active, drawing on insights from both philosophy and cognitive science. Three cases will be central: mind-wandering, cognitive maps, and predictive processing. The dissertation begins with an introduction outlining the guiding research objections and the current state of relevant fields of research. The main body consists of four articles, each examining different aspects of thought. The first article develops an account of the role of mind-wandering in planning that reconciles

its exploratory function with the need to maintain stable intentions for long-term planning. The second article challenges the view that mind-wandering is passive and unguided, arguing instead that it is an actively guided learning process that yields conclusions about what is the case or what to do. The third article argues that mental transitions mediated by cognitive maps qualify as reasoning, even though they lack the rule-governed operations over propositional attitudes often thought to define reasoning. The fourth analyzes two strands of the predictive processing framework: one that collapses the distinction between beliefs and desires, and another that reintroduces it. The analysis shows that maintaining this distinction is indispensable for a complete account of agency and the mind. The dissertation concludes by discussing how the results of these investigations bear on broader questions about the representational nature of the mind.

Research Objectives

Debates concerning the nature of thought have often centered around what makes our thinking rational and active. One aspect of this concerns the forms of awareness we do or do not have of what transpires in thought. Another concerns what kind of cognitive architecture underpins our thinking. This architecture can be examined from several angles: the types of content it admits, the formats in which this content is represented, the operations it enables, and the ways in which mental attitudes such as beliefs, desires, and intentions are instantiated. Against this background, the dissertation addresses the following general research questions:

- What forms does agency take in thought?
- What forms of rationality are exhibited by thought?
- What forms of awareness do we have of our own thinking?
- What are the contents and formats of representations in thought?
- What types of operations occur in thought?
- What is the nature and role of mental attitudes in thought and action?

Philosophers have long grappled with these questions, and a vast body of work has explored them in depth. This dissertation adopts a more piecemeal approach, focusing on what specific mental phenomena reveal about the nature of thought. In doing so, it moves beyond the paradigmatic cases of self-aware, active, and logically structured thought to investigate forms of thinking that fall outside the traditional focus of philosophical inquiry, while drawing on insights from cognitive science into their underlying mechanisms.

The dissertation centers on three cases that seem to challenge conventional wisdom: mind-wandering, cognitive maps, and predictive processing. Mind-wandering, often regarded as passive and purposeless, is typically contrasted with rational and active processes such as reasoning, planning, and executive control. Yet emerging work suggests it plays functional roles in planning and learning, overlapping with processes usually seen as rational and active.

Cognitive maps—mental representations of geometric structure—seem to enable agents to reach conclusions about what is the case or what to do in rational ways, despite (often) lacking logical and propositional structure. This raises the possibility that some reasoning is mediated by non-logical, non-propositional representational structures, challenging the traditional view of reasoning as rule-governed operations over propositional attitudes.

Finally, proponents of the recently popular predictive processing framework have suggested that we ought to dispense with traditional distinctions between beliefs and desires, raising questions about the role of these attitudes in thought and action. To advance our understanding of thought, the dissertation therefore pursues the following more specific research questions:

- What is the role of mind-wandering in planning, and how does it relate to theories of rational planning agency?
- What are the mechanisms underlying mind-wandering, and what do they reveal about its connection to rational and agential capacities?
- What role do cognitive maps play in reasoning, and what implications does this have for our understanding of the nature of reasoning?
- Should we revise traditional distinctions between beliefs and desires as indicated by certain predictive processing theories, or are these categories indispensable?

Through these inquiries, the dissertation develops novel accounts of core aspects of our mental lives, rethinks the boundaries of rational and active thought, and sheds light on the representational and computational mechanisms underlying diverse forms of thought.

State of the Art

In this section, I review current understandings of thought, action, reasoning, and the kinds of cognitive architecture that may support these. In the process, I highlight gaps in existing accounts that this dissertation seeks to address.

Action

The issue of agency in thought is connected to the more general problem of action. In philosophy, the problem of action concerns what it is for an event or process to count as something an agent does, rather than something that merely happens. This distinction—between genuine actions and mere happenings—raises deep questions about the conditions under which we can truly say of someone that they acted. Philosophers have long sought to explain what marks out actions from occurrences beyond an agent's control, and to articulate the role of reasons, intentions, and various kinds of control in that explanation.

At the heart of the problem is a set of interconnected puzzles. How should we characterize the relation between an agent's mental states—beliefs, desires, intentions—and the physical happenings that constitute their behavior? Can an action be fully explained in terms of its causally antecedent mental states and their role in the rational organization of an agent's life? Is a special kind of knowledge of what one is doing required for action? These questions are pressing in the domain of bodily action, where philosophical reflection has generated an intricate set of debates about what is required for an event or process to qualify as an action.

The dominant notion of action in philosophy remains that of intentional action. Following Donald Davidson's (1963) seminal formulation, intentional actions are typically given both a causal and a rationalizing explanation: intentional actions are events caused by the agent's mental states—beliefs, desires, intentions—and can be rationalized by citing the agent's reasons—beliefs and desires—that explain why the agent acted as they did. For example, you raised your arm because you wanted to get home quickly and believed that such signaling would make a taxi stop for you. Your action makes sense by reference to your reasons—your desire to get home and your belief about how to achieve it. These mental states did not merely rationalize the action; they caused it. The belief–desire pair explains why the movement of your arm counts as an intentional action (hailing a cab) rather than a mere bodily event (e.g., your arm rising due to a muscle spasm). This account, known as the Causal Theory of Action, emphasizes the causal antecedents of action and, given its continued influence, is often referred to as the standard story of action (Piñeros Glasscock & Tenenbaum, 2023).

Another influential approach, stemming from the work of Michael Bratman, treats action not merely as a punctate event but as something that can extend over time. For Bratman (1987), extended actions are unified by the agent's capacity for planning: a form of practical reasoning that structures and guides our activities over time. Planning involves settling on and committing to courses of action in advance. This process forms future-directed, partial, and hierarchically

structured intentions. These intentions allow agents to coordinate their activities over time and with others, while managing limited cognitive resources by reducing the need for constant deliberation. Rather than deciding how to act in the moment, we often execute prior intentions formed earlier. Future-directed intentions are initially partial: we commit to long-term goals without yet specifying all the steps needed to achieve them. Through subsequent episodes of practical reasoning, we gradually fill in these details and come to intend the requisite means to our long-term ends. To fulfill these roles, future-directed intentions need to remain stable over time and resist frequent reconsideration. When our intentions remain appropriately stable, coherent, and consistent with one another and with our beliefs, we qualify as rational planning agents. Extended action, on this view, is not a mere sequence of momentary doings but a unified whole, held together by the structuring role of intentions and the rational pressures toward stability, coherence, and consistency.

To illustrate, suppose you are deciding how to spend your upcoming summer: whether to take a relaxing beach holiday or embark on a hiking trip. You eventually settle on one option—say the hiking trip—thereby resolving the question and ending the need for further deliberation on this matter. This intention then structures subsequent reasoning about what to do: you research destinations, book accommodations, purchase equipment, and adjust your schedule to prepare for the trip. As your plans unfold, you form more concrete intentions that guide each step. For this to work, your intentions must remain fairly stable over time, you must adopt means that cohere with your ends, and your intentions must be consistent with your other intentions and beliefs; otherwise, they cannot effectively structure extended action.

Another influential line of thought emphasizes that action is not fully explained by its causally antecedent mental states but is essentially a matter of being guided by the agent as it unfolds. Harry Frankfurt (1978) famously argued that what distinguishes an action from a mere bodily movement is not simply that it is caused by an intention or belief-desire pair, but that it is subject to the agent's guidance and control while it is occurring. This dynamic element—monitoring and regulating what one is doing in the course of doing it—marks out genuine action from movements that happen to be caused by a relevant intention but fail to remain under the agent's ongoing guidance. Action, on this view, is thus not merely a matter of initiating a causal chain of events but of exercising control over an activity as it unfolds.

Finally, empirically informed philosophers have drawn on cognitive science to develop accounts of the mechanisms that enable agents to select, initiate, and control their actions. Proposals highlight coordinated hierarchies of intentions and motor representations (Pacherie, 2008; Mylopoulos & Pacherie, 2019), the role of attention in selecting features to guide action

(Wu, 2016), the role of executive control mechanisms in guiding action (Pacherie & Mylopoulos, 2020; Buehler, 2022; Shepherd, 2025), and the role of value representations in action selection (Carruthers, 2025; Sripada, 2025).

Although these perspectives in some respects offer competing accounts of action, each has yielded important insights into the nature of action. Building on these insights, and integrating findings from cognitive science, this dissertation argues that many of the mechanisms that guide bodily action also guide the way our thinking unfolds.

Mind-wandering

A closer examination of agency in thought involves identifying which forms of thought deploy mechanisms constitutive of action. This may involve revising what common-sense recognizes as active thinking. To probe the extent of mental agency, it is instructive to consider borderline cases. A prominent example is mind-wandering, a form of thinking often regarded as passive. Research on mind-wandering has recently developed into a vibrant interdisciplinary field at the intersection of philosophy, psychology, and cognitive neuroscience.

Contemporary research has sought to provide a systematic characterization of mind-wandering, including its underlying mechanisms, functions, and agentive status, yet there remains disagreement about how to define it. Mind-wandering has been variously described as a shift of attention away from a task or perceptual input toward self-generated content (Smallwood & Schooler, 2015), as sometimes intentional and sometimes unintentional (Seli et al., 2016, 2018), and as unguided thought that meanders freely between loosely related topics (Christoff et al., 2016; Irving, 2016, 2021). Some argue that the construct is heterogeneous, encompassing different combinations of multiple overlapping features (Seli, 2018), while others resist this pluralism and defend a unitary definition (Christoff et al., 2018).

Empirical research has largely operationalized mind-wandering as episodes in which participants report that their thoughts have drifted from an experimental task to task-unrelated content (Smallwood & Schooler, 2015). This ‘task-unrelated thought’ paradigm has anchored a large body of research on the frequency, content, functions, and neural and cognitive mechanisms of mind-wandering. Studies show that mind-wandering occupies a large portion of waking life. Estimates suggest we mind-wander between 30-50% of our waking hours (Kane et al., 2007; Killingsworth & Gilbert, 2010). Mind-wandering has also been associated with important cognitive functions. Its content often relates to the future, ourselves, and our goals, suggesting a role in planning (Baird et al., 2011; Stawarczyk et al., 2011, 2013). Mind-

wandering has also been found to improve creative problem-solving (Baird et al., 2012; Gable et al., 2019). Neuroscientific studies associate mind-wandering with activity in large-scale brain networks, including the default mode network—active during rest, self-referential thought, and episodic simulation—and executive control and salience networks (Christoff et al. 2009; Fox et al., 2015; Turnbull et al., 2019).

These findings have driven recent philosophical interest in mind-wandering, though interpretations vary significantly. Philosophical work has largely centered on its implications for mental agency. Emphasizing mind-wandering's tendency to deviate from ongoing tasks or goals, some philosophers argue that it lacks the monitoring and regulation required to actively guide attention or thought toward one's goals (Irving, 2016, 2021; Murray, 2025). The presumed absence of such monitoring and regulation has led these philosophers to conclude that mind-wandering is a passive, unguided phenomenon.

Other theorists, often focused on explaining the functions of mind-wandering, have offered more agency-oriented accounts of mind-wandering. Some propose that mind-wandering is a form of mental exploration, searching for new and potentially better goals or opportunities for action to pursue (Sripada, 2018; Shepherd, 2019). These accounts often emphasize how cost-benefit computations sometimes favor exploration, triggering mind-wandering when it is deemed more valuable than maintaining focus on the current task. Joshua Shepherd (2019) suggests that this value-based switch is implemented by the executive control system, indicating that mind-wandering is in fact guided by the agent's goals and values. Similarly, Peter Carruthers (2015) argues that associatively activated representations are monitored for their relevance to one's goals and values during mind-wandering. When deemed relevant, they draw attention and may enter working memory, where they can be maintained and manipulated until replaced by new representations. This, Carruthers contends, indicates that the content of mind-wandering is actively selected for further processing by executive control networks.

Despite these advances, much remains unknown about the workings of mind-wandering. A promising avenue of research is to clarify how it supports the functions attributed to it. Identifying the mechanisms through which mind-wandering fulfills these functions may help resolve whether it is genuinely passive or actively guided. Connections with research on planning and executive control are especially intriguing, as both involve distinct forms of agency and appear to relate to mind-wandering in ways that are not yet fully understood.

Examining the relationship between mind-wandering and rational planning agency may shed new light on its role in planning. If mind-wandering explores new and potentially better goals and actions, it may frequently provide reasons to reconsider existing intentions in favor

of better alternatives. At first glance, this appears to threaten the stability of intentions needed to achieve long-term goal, suggesting a potential conflict with rational planning agency. Yet mind-wandering may also contribute to planning in ways that are compatible with rational planning agency. Article 1 develops a proposal for how this might occur.

A closer inspection of the relationship between mind-wandering and executive control promises to help determine whether mind-wandering is passive or active. Executive control comprises capacities of individuals to monitor and regulate their thoughts and actions in pursuit of goals. These include capacities to activate goal-relevant representations, maintain and manipulate these in working memory, enhance goal-relevant processing, and inhibit distractions and prepotent responses. Exercising these capacities is taken to constitute active guidance of one's activities (Buehler, 2022). Because mind-wandering is often described as lacking such control capacities, it is often regarded as a passive, unguided phenomenon (Irving, 2016, 2021; Murray, 2025). This raises a puzzle: if mind-wandering lacks the mechanisms that guide one's activities toward one's goals, how can it contribute to goal pursuit? Article 2 addresses this question.

Reasoning

Another form of thinking more commonly associated with rationality and agency is reasoning. Philosophers typically distinguish between theoretical reasoning, which concerns what is the case or what to believe, and practical reasoning, which concerns what to do or what to intend. Beyond this general distinction, theories differ substantially on what reasoning consists in and how these two general forms relate.

In one respect, reasoning seems to contrast with mind-wandering, as the latter is commonly conceived: reasoning is something we actively do. It is often defined as a conscious, intentional mental activity in which an agent actively constructs, evaluates, and regulates a series of inferential steps to reach a conclusion. The reasoner is not a mere passive observer of their thoughts but an active participant, directing their thinking to shape their understanding of a subject matter and arrive at well-founded conclusions.

Beyond being an active transition from premise-states to a conclusion-state, a prominent strand within philosophy also takes reasoning to involve a specific kind of operation over particular states. Reasoning, on this view, is a rule-governed operation over propositional attitudes (or their contents) (Broome, 2013; Boghossian, 2014). John Broome (2013) summarizes the view succinctly:

> Active reasoning is a particular sort of process by which conscious premise-attitudes cause you to acquire a conclusion-attitude. The process is that you operate on the contents of your premise-attitudes following a rule, to construct the conclusion, which is the content of a new attitude of yours that you acquire in the process (p. 234).

The rules in question concern how to maintain consistency and coherence among one's propositional attitudes. They are broadly logical, employing logical operators such as NOT, AND, and IF-THEN. While these rules are not limited to deductive reasoning, inference rules from deductive logic are paradigmatic examples. A classic instance is the modus ponens rule:

> If you believe that *if P, then Q*, and you believe *P*, then you ought to believe *Q*

Although practical reasoning is not deductively valid—it is non-truth-preserving, since it operates not only over truth-apt beliefs but also over non-truth-apt intentions and desires—it still aims to maintain coherence and consistency among one's beliefs, desires, and intentions. Still, the rules governing practical reasoning also feature logical terms such as IF-THEN.[1] Consider, for example, the means-end coherence rule:

> If you intend an end, and you believe that certain means are necessary to achieve it, you ought also to intend those means.

These are the core commitments of the standard rule-following account of reasoning, though the view admits of several possible extensions. One important question concerns the format in which reasoning occurs. Broome, for example, considers it plausible that reasoning must be made explicit and proposes that language provides the natural medium for this:

> In active reasoning, you operate on the marked contents of your conscious attitudes, following a rule. These marked contents are complex. They have a syntactic structure, and the rules you apply in operating on them depend on their structure. In operating on them, you have to hold them in your consciousness, maintaining an awareness of their syntactic structure. Language is well suited to doing that. It has a meaning that can represent the

[1] Further issues arise in the formulation of rules, including whether rules of reasoning require that you perform the inferences prescribed or merely permit it. It is also debated whether, for certain rules, the 'ought operator' has narrow or wide scope, that is, whether it applies to the consequent of a conditional (if you believe you ought to do something, you ought to do it) or to the entire conditional (you ought that, if you believe you ought to do something, you do it). Narrow-scope formulations require a single response to ensure coherence between mental attitudes, whereas wide-scope formulations permit multiple. For example, a narrow-scope formulation requires that you form the intention to do something if you believe you ought to do it, whereas a wide-scope formulation also permits you to give up the belief that you ought to do it. Since these issues will not be important for my purposes, I will gloss over them.

> semantic elements of the marked contents, and it has a syntax that can represent their syntactic structure. It is plausible that, without the help of language, you could not keep the marked contents (ibid., p. 267).

Marked content consists of a proposition along with an indication of the attitude taken toward it. Broome claims that reasoning operates over such marked contents, taking them as inputs and yielding new marked contents as outputs. For example, the belief that *I shall take a break* will have the marked content <I shall take a break; belief>, distinguishing it from, say, an intention with the same proposition as content. This is important, since the output of correct reasoning is not just an unmarked proposition but a specific attitude toward that proposition. Consider, for example, Broome's formulation of a rule for correct instrumental reasoning:

> From
>
> < *E* ; intention > and
>
> < *M* is a means implied by *E* ; belief > and
>
> < *M* is up to me; belief >
>
> to derive
>
> < *M* ; intention >

This rule specifies that correct instrumental reasoning takes as input: (i) an intention to pursue some end, (ii) a belief that the end requires producing a certain means, and (iii) a belief that it is up to me whether the means are produced—that is, if I were not now to Intend *M*, because of that *M* would not come. From these inputs, the reasoning process yields as output an intention to pursue the means to the end. This process can be made explicit via language, as illustrated by the following example:

> 'I shall visit Venice
>
> My buying a ticket is a means implied by my visiting Venice
>
> My buying a ticket is up to me
>
> So, I shall buy a ticket'

But expressed this way, the reasoning could be confused with theoretical reasoning. The first premise-sentence, 'I shall visit Venice', could be interpreted either as a belief that I shall visit Venice or as an intention to do so. If interpreted as a belief, I could correctly derive the belief that I shall buy a ticket from the premises. But this would constitute theoretical reasoning, which takes beliefs as input and outputs a further belief, rather than instrumental reasoning,

which takes an intention and certain beliefs as input and outputs an intention. The line of reasoning can instead be made explicit in the following way:

'I intend to visit Venice
My buying a ticket is a means implied by my visiting Venice
My buying a ticket is up to me
So, I shall buy a ticket'

Expressed this way, the first premise-sentence is unambiguously marked as an intention. The line of reasoning is now a plausible example of correct instrumental reasoning, no longer to be confused with theoretical reasoning.[2]

If reasoning occurs in language, it is natural to suppose that the underlying cognitive architecture resembles a language of thought, with a combinatorial syntax and semantics. The language of thought hypothesis (Fodor, 1987; Fodor & Pylyshyn, 1988) holds that mental representations are structured like sentences, composed of symbols that combine according to formal rules to express complex thoughts. This structure supports productivity and systematicity of thought. Productivity is the capacity to entertain a potentially unlimited number of thoughts from a finite set of constituents, constrained by biological factors such as memory, attention, processing capacity, etc. For example, using the same constituents, I can think that *she was my mother; she was my mother's mother; she was my mother's mother's mother,* and so on. Systematicity means that the capacity to entertain some thoughts is inherently connected to the capacity to entertain structurally related thoughts with the same constituents rearranged. For example, if you can entertain the thought that *John loves Mary*, then you can also entertain the thought that *Mary loves John*.

This view is associated with a computational view of the mind, in which cognitive processes are understood as symbolic manipulations over these complex, logically structured representations. Building on this view, Jake Quilty-Dunn and Eric Mandelbaum (2018) argue that the rule-governed operations characteristic of reasoning are enabled by a cognitive architecture in which logical rules—such as modus ponens—are built in, and these rules operate over language-like representations. Linking reasoning to language-like representations thus potentially has important implications for what cognitive architecture might support it.

[2] Note that Broome does not think that the conclusion can be reformulated as the sentence, 'So, I intend to buy a ticket'. He interprets this as incorrect theoretical reasoning, since it appears to conclude in the belief that one intends to buy a ticket. Yet even if one holds this intention, it does not follow that one believes that one holds it.

In addition, it is often argued that reasoning requires a thinker to appreciate, in some sense, that the conclusion is supported by the premises, basing their conclusion on their premises for this reason. Paul Boghossian (2014) articulates this idea through the following condition:

> *Taking Condition*: inferring necessarily involves the thinker *taking* their premises to support their conclusion and drawing their conclusion *because* of that fact (p. 5; original emphasis).

According to Boghossian, taking is constituted by the act of applying a rule. If you apply modus ponens to *P* and *if P, then Q* to derive *Q*, the very act of following the rule involves, in some sense, taking the premises to support the conclusion. This process is supposed to be conscious—or at least potentially conscious:

> 'reasoning [is] a mental action that a person performs, in which he is either aware, or can become aware, of why he is moving from some beliefs to others' (ibid., p. 16).

These conditions are meant to differentiate genuine inferential transitions from purely causal ones, where mental states are produced by causal processes that potentially mimic correct reasoning without involving actual reasoning. Several theorists, however, maintain that processes can count as reasoning even if they lack features emphasized by rule-following accounts. It has been argued that inference need not involve conscious taking (Quilty-Dunn & Mandelbaum, 2018; Siegel, 2019; Levy, 2024) or be made explicit in linguistic or logical form (Buckner, 2019; Munroe, 2021; Levy, 2024; Shea, 2024a, b).

This work suggests that there are alternate ways to mark out genuinely inferential transitions. Many extant accounts share that reasoning involves drawing conclusions in response to premise-states in ways that are responsive to rational norms. As we will see, plausible candidates for meeting these conditions involve responding to the content of preceding states in content-preserving ways and monitoring and regulating cognitive strategies according to their quality and cost.

A few important distinctions deserve clarification at this point. The term *inference* is often used interchangeably with *reasoning* (Broome, 2013; Boghossian, 2014; Buckner, 2019; Siegel, 2019; Munroe, 2021), and I will follow this convention. Some reserve *reasoning* for step-by-step, rule-governed transitions between thoughts, while using *inference* more broadly to include transitions based on representational content beyond that of purely logical terms. In rule-governed transitions such as modus ponens, the non-logical terms (*P* and *Q*) can be freely substituted, and the transition remains truth-preserving. By contrast, transitions such as moving

from the thought *Cyrus is a dog* to the thought *Cyrus barks* are not guaranteed to preserve truth. Still, their shared (dog-related) content means that the latter thought is likely to be true if the former is. A thinker might be disposed to make this transition because of the observed regularity that *dogs bark*, without explicitly representing that fact, say, in conditional form: *if X is a dog, then X barks*. Following Nicholas Shea (2024a), we can think of this as a content-specific transition: whether the transition is faithful to content depends on the content of non-logical terms. The distinction between rule-governed (or content-general) transitions and content-specific transitions will become relevant in the discussion of map-based reasoning.

Another important distinction concerns deliberative versus non-deliberative attitude change. Deliberation is initiated by an intention to settle an issue and reach a conclusion, whereas non-deliberative change is not intentional in this way. The latter is needed to avoid regress: if every attitude change required deliberation, then each act of deliberation would itself require prior deliberation to form the intention that initiates subsequent deliberation. Non-deliberative attitude change (particularly, intention formation) provides the starting points that make deliberation possible (Mele, 2003, Chap. 9; Arpaly & Schroeder, 2012; Railton, 2017). This distinction will be key in the discussion of attitude change during mind-wandering.

Cognitive maps

One way to challenge the assumption that reasoning operates exclusively over linguistic, logical, or propositional representations is to show that other representational formats also support transitions that qualify as reasoning. One plausible candidate is cognitive maps. A cognitive map is an internal representation of geometric relations, which include metric relations (e.g., distances and angles) and topological relations (e.g., connectedness and adjacency). Understood in a stricter sense, cognitive maps themselves exhibit geometric structure and are veridical (or accurate) when this structure corresponds to that of the environment (Rescorla, 2009a, 2017). Cognitive maps are often contrasted with language-like representations, because they (typically) lack mechanisms that can play the role of predicates, logical operators, or quantifiers, which give language-like representations their logical and propositional structure (Camp, 2007, 2018; Rescorla, 2009a, 2009b).

Cognitive maps have been most extensively studied in the context of spatial navigation. One key function they support is route planning: an agent constructs potential paths between a starting point and a destination and selects the most efficient route by representing locations

and their connecting paths on a cognitive map. Because planning can be a form of practical reasoning (Bratman, 1987), cognitive maps may contribute to this type of reasoning.

Recent evidence indicates that cognitive maps can also encode more abstract relationships. One study found that participants created a two-dimensional map of a social hierarchy, with individuals positioned according to competence and popularity. This map allowed participants to identify pairs of individuals who achieved the best balance between the two traits (Park et al., 2021). Superficially, this resembles a reasoning process: participants transform representations of individuals and their traits into conclusions about optimal partnerships.

Philosophical and computational work on cognitive maps also suggests that map components can be flexibly recombined to represent novel geometric structures. For example, a blue blob in one quadrant of two crossing lines can represent a lake at a road junction, while the same blob between parallel lines can represent a lake between parallel roads (Camp, 2007).

When constructing cognitive maps, the brain appears to separate representations of entities from the structures they inhabit, which enables flexible recombination of map components and allows knowledge from one situation to generalize to structurally similar ones. Among the components that make up a cognitive map are different types of cells that represent distances and directions to objects, borders, landmarks, or goals (Whittington, 2020, 2022). This compositional flexibility makes cognitive maps useful for reasoning about relationships that can be represented geometrically. Some tasks—such as determining geometric relations among multiple entities—are more efficiently solved by locating entities on a map than by listing geometric facts in sentences and deducing additional relations from previously listed facts. This suggests that certain forms of reasoning are better served by cognitive maps.

Different representational formats afford different kinds of computation, so broadening the range of formats admissible in reasoning correspondingly expands the kinds of computations that might be used in reasoning. Cognitive maps are naturally suited for computing directions and distances. Yet if cognitive maps lack logical and propositional structure, and reasoning is defined as rule-governed operations over propositional attitudes (or their contents), it is unclear how maps could support reasoning. Rule-governed reasoning is naturally implemented by language-like representations, which provide the logical and propositional structure required for such operations. Maps (typically) lack such structure.

This gives rise to a tension between the idea that cognitive maps can support reasoning and the rule-following account. We can address this tension in several ways: (1) deny that reasoning operates over cognitive maps, (2) attempt to reconcile map-based reasoning with the rule-

following account, or (3) adopt an alternative account of reasoning that treats transitions other than rule-governed ones as inferential. These options are explored in Article 3.

Cognitive architecture

The question of what kinds of representations we employ in thought naturally raises further questions about the cognitive architecture that supports the capacities of thought. Because representational structures constrain the types of computation that can be performed over them, identifying the structures involved in thought also clarifies what forms of computations we perform in thought. This link is especially clear in the language of thought hypothesis (Fodor, 1987; Fodor & Pylyshyn, 1988; for a comprehensive overview, see Rescorla, 2023):

1. *Mental representations have a combinatorial syntax and semantics*. Complex representations are composed of simpler constituents and the meaning of complex representations depends on the meaning of their constituents as well as the constituency structure into which the constituents are arranged. Mental representations have logical structure such that their compositional semantics resemble that of logically structured linguistic expressions. For example, when you believe that *P and Q*, what you believe is composed of the proposition that *P* and the proposition that *Q*.
2. *The mental processes that operate over mental representations are sensitive to the structure of those representations*. Computations resemble those of a Turing machine, operating over discrete mental symbols according to formal rules. For example, during an inference one might apply an operation to a representation of the form *P and Q* to transform it into a representation of the form *P* in a process of conjunction elimination.

In this cognitive architecture, the logical structure of mental representations naturally lends itself to logical operations over those representations. In other words, representational structure begets certain forms of computation. However, over the decades it has become common to acknowledge a wider variety of representational structures and, correspondingly, a wider variety of computational processes in the mind. In recent years, different computational approaches, including Bayesian learning, reinforcement learning, and neuroeconomics, have identified a wide range of computational processes in the mind beyond symbol manipulation.

Recent proponents of the language of thought hypothesis also acknowledge a plurality of representational structures and computations (Quilty-Dunn et al., 2023).[3]

Another core aspect of cognitive architecture concerns the treatment of mental attitudes. In standard philosophical accounts, beliefs and desires play central and distinct roles in explaining thought and action. Many computational theories of the mind are consistent with this. A common way of distinguishing beliefs and desires is through the notion of direction of fit. Desires have a world-to-mind direction of fit: in desiring something, we aim to make the world fit our desire. Beliefs, by contrast, have a mind-to-world direction of fit: we aim to fit our beliefs to how the world actually is.

This distinction points to the need for distinct mechanisms to update belief-like and desire-like states. This aligns with the Humean theory of motivation, according to which beliefs and desires are fundamentally distinct mental states, and desires play a necessary role in motivating action (Smith, 1987)—a view widely assumed in philosophy and cognitive science. The Humean position is evident in influential computational frameworks such as Bayesian decision theory and reinforcement learning.

In Bayesian decision theory, belief-like states are captured by probability assignments reflecting degrees of belief in hypotheses, while desire-like states are captured by a utility function representing preferences over outcomes. Possible outcomes are assigned values, reflecting their desirability, and weighted by the probability of their occurrence; actions with the highest expected value are most likely to be chosen. Similarly, in reinforcement learning values are assigned to actions or outcomes, probabilities are assigned to reaching certain outcomes given certain actions, and action selection favors actions with the highest expected value. In both frameworks, desires (value assignments) provide the motivational force driving action, while beliefs (probability assignments) inform decision-making by representing the structure of the environment and likely consequences of actions. The separation of belief-like and desire-like states, and their coordinated role in action selection, reflects the core commitments of the Humean theory.

Predictive processing

More recently, predictive processing has emerged as an influential framework in philosophy and cognitive science (Hohwy, 2013; Clark, 2016; Parr et al., 2022). Predictive processing

[3] See also Goodman et al. (2015) on how to accommodate probabilistic inference within a language of thought architecture.

offers a computational framework that diverges from the traditional Humean view. Mental processes are modelled as hierarchical prediction and prediction-error minimization. Perception involves generating predictions about sensory inputs and updating them in light of prediction errors, while action involves predicting sensory consequences and acting to fulfill those predictions.

A major debate concerns whether predictive processing can adequately account for motivation and action. Critics contend that a distinct desire-like primitive with a world-to-mind direction of fit is required to explain what motivates action (Colombo, 2017; Klein, 2018, 2020). Traditional formulations of predictive processing posit only one primitive—prediction—expected to serve both belief-like (fitting the world) and desire-like (shaping the world) roles (Clark, 2020). However, this dual role creates tension: belief-like predictions must be revisable to continuously fit the world, whereas desire-like predictions must resist revision to drive action. Without a clear distinction between the two, maladaptive outcomes may result in contexts that demand independent updating of beliefs and desires pertaining to the same state of affairs—unlike for Humean agents, who separate these states and thereby allow independent updates (Klein, 2020).

Recent predictive processing models introduce expected free energy minimization, selecting actions by balancing uncertainty reduction against progress toward preferred outcomes (Parr et al., 2022). By incorporating progress toward preferred outcomes as an independent concern, these models may restore a distinction between belief-like and desire-like states, aligning them more closely with Humean frameworks—but at the cost of the simplicity of earlier predictive processing models. These issues are explored in Article 4.

Methodology

The central aim of this dissertation is to clarify the nature of rational and active thought by examining both the conditions that make such thought possible and the forms of thinking that satisfy those conditions. The project adopts a naturalistic approach: philosophical theories (of the mind) should be informed by the best available science. Cognitive science offers crucial insights into the workings of the mind, and it would be a missed opportunity not to integrate these findings. This commitment does not reduce philosophy to data analysis; rather, it encourages models that are both conceptually coherent and scientifically plausible. Attention is paid to the assumptions embedded in experimental paradigms and scientific models, and empirical cases are treated as opportunities for philosophical investigation. In some cases, this

prompts a re-examination of traditional theories when they conflict with what science reveals about the phenomena under study.

Phenomena should be subjected to careful philosophical analysis of key concepts and arguments and interpreted for their broader significance to questions of philosophical—and, hopefully, general—interest. Guided by empirical research, the project investigates the mechanisms underlying key mental phenomena and uses these insights to inform and refine philosophical theorizing. Although we should be cautious not to give too much weight to intuitions evoked by thought experiments when empirical evidence is lacking or in conflict, imagined scenarios remain valuable for clarifying key points, exposing assumptions, and probing the implications, coherence, and limits of theories.

Although continuous with science in many respects, this approach remains recognizably philosophical in its more abstract aims and argumentative style. Philosophers are trained to clarify concepts and distinctions, construct and analyze arguments, think critically and systematically about implications and objections, and connect issues to fundamental questions. Philosophical analysis helps refine concepts, capture phenomena with precision, and integrate results into coherent, overarching theories. Freed (mostly) from the demands of running laboratories, philosophers have time to read widely, synthesize insights across diverse disciplines that are rarely brought into contact, and frame results in terms of foundational questions. This involves immersing oneself in scientific literature and extracting lessons that bear on philosophical problems, while abstracting away from technical minutiae that occupy specialized scientists to discern broader patterns across phenomena often studied in isolation. Conversely, the philosophical literature can be a source of new hypotheses, research questions, and frameworks that fruitfully inform and help interpret scientific inquiry.

The dissertation proceeds through a series of case studies, each raising questions about thought, rationality, agency, and cognitive architecture. These cases test the limits of traditional theories and help us explore the merits of alternative ones. The overarching goal is to provide a better understanding of key aspects of the mind that is attentive both to philosophical rigor and empirical evidence.

Article Previews

The dissertation comprises four articles, each addressing distinct but interconnected aspects of rationality and agency in thought. In this section, I summarize the central arguments and conclusions of each article.

Article 1:

Is the Wandering Mind a Planning Mind?

The first article investigates the role of mind-wandering in goal exploration and planning. While empirical research supports the idea that mind-wandering serves these functions, their precise mechanisms remain poorly understood. The article raises a problem with the view that the sole function of mind-wandering is to explore goals that are potentially better than the ones we are currently committed to: extensive exploratory mind-wandering could lead agents to frequently find reason to reconsider their intentions, thereby threatening the stability required for rational planning agency.

In response, the article offers a model that integrates the exploratory role of mind-wandering with the stability of intentions essential for rational planning agency. It is argued that, beyond exploring new goals, mind-wandering supports other planning-related functions: identifying means to achieve one's ends and identifying reasons that favor existing intentions. Rather than undermining intentions, these functions help stabilize them. Thus, although mind-wandering may at times generate reasons to reconsider and prompt reconsideration, this is counterbalanced by other functions that sustain commitments to long-term goals.

Changes to one's beliefs are a core way in which what one has reason to do might change. The article therefore explores the implications of belief changes brought about by mind-wandering. Combining a Spinozan model of belief formation—where every entertained proposition is automatically believed—with a fragmentation model of belief storage—where beliefs are stored in independent fragments—has implausible consequences if mind-wandering can alter one's beliefs. Because the wandering mind entertains a vast array of propositions, these models together imply that mind-wandering could produce a rapid accumulation of inconsistent beliefs across fragments. Such a belief set would complicate coordinating actions over time, undermining rational planning agency. Given that we often act as rational planning agents, we have reason to reject Spinozan fragmentation model hybrids.

Finally, the paper argues that mind-wandering presents a plausible candidate for non-deliberative attitude change. Mind-wandering is not initiated and guided by intentions to deliberate, and resulting changes in attitudes therefore count as non-deliberative.

Article 2:

Mind-Wandering in Action

The second article argues that explaining how mind-wandering contributes to goal pursuit requires rethinking its status as passive and unguided. Although mind-wandering is often described as passive, it also serves important goal-related functions. Attempts to resolve this tension typically portray mind-wandering as at once passive and purposive. This article challenges that compromise, arguing that a satisfactory account of mind-wandering's role in goal pursuit rules out passivity. Drawing on empirical evidence and philosophical analysis, it is shown that mind-wandering contributes to goal pursuit through mechanisms that monitor, evaluate, and regulate mental representations and processes, some of which amount to active guidance by the executive control system. This contradicts the prevailing view of mind-wandering as passive and unguided. Instead, mind-wandering is an actively guided learning process that enables us to reach conclusions about what is the case or what to do.

This has significant implications for our understanding of rational inference. Since some of the processes occurring during mind-wandering—imagining suppositional scenarios, evaluating them against our goals and values, and drawing conclusions on this basis—often support rational inference, there is some reason to think that conclusions reached during mind-wandering can qualify as rational inference.

The account also has important implications for mental agency. During mind-wandering, different features of agency appear to come apart. We are typically unaware that our minds have wandered and of what might have triggered it, indicating a lack of knowledge of what one is doing that is sometimes associated with agency. Moreover, mind-wandering does not seem to be guided by intentions aimed at settling an issue and reaching a conclusion, which guide deliberative thought. The effects of mind-wandering on our attitudes are therefore better classified as non-deliberative attitude change. If mind-wandering is actively guided by our goals and values and leads us to draw conclusions about what is the case and how to act, this might also imply that we are in some ways responsible for how we mind-wander.

Article 3:

Reasoning with Cognitive Maps

The third article makes the case that cognitive maps can facilitate reasoning. It identifies a tension between three claims: 1) reasoning is a rule-governed operation over propositional attitudes (or their contents); 2) reasoning can operate over cognitive maps; and 3) cognitive

maps lack logical and propositional structure. It is argued that we should deny the first claim and opt for a more inclusive account of reasoning. Mental transitions mediated by cognitive maps can meet plausible conditions for reasoning: conclusions are responses to premise-states; transitions are responsive to rational norms; and the reasoner takes their conclusion to be supported by preceding states and operations.

It is argued that cognitive maps can mediate content-specific transitions from questioning attitudes to conclusion-states that respond with relevant answers. Questioning attitudes motivate mental searches for answers and, as part of this search, cognitive maps help structure mental simulations that identify relevant answers. These transitions are responsive to certain rational norms: they optimize for reliability and expected value. Metacognitive processes monitor the quality and costs of map-mediated transitions and regulate their use, making it likely that such transitions are deployed in tasks for which they are especially well-suited. When a strategy has proven favorable, it will be accompanied by metacognitive feelings of fluency, control, reliability, or confidence. Such feelings likely accompany map-mediated transitions, and when they do, the reasoner can plausibly be said to take their conclusion to be supported by the preceding states and operations. Since map-mediated transitions are not rule-governed operations over propositional attitudes yet still satisfy plausible conditions for reasoning, they challenge the rule-following account.

Article 4:
Predictive Minds Can Be Humean Minds

The third article contends that the predictive processing literature encompasses two distinct versions of the framework. One version—dubbed optimistic predictive processing—uses the notion of optimistic priors to account for agents' motivation to act. A more recent iteration—dubbed preference predictive processing—explains action selection as minimization of expected free energy. Despite offering vastly different accounts of motivation and action, these two approaches are often conflated in the literature.

Optimistic predictive processing reduces belief-like and desire-like states to a single construct—prediction. This challenges standard philosophical and scientific accounts of motivation and agency, which assume a clear distinction between belief-like and desire-like states and attribute a necessary motivational role to desire-like states—a thesis known as the Humean theory of motivation. By contrast, preference predictive processing introduces distinct desire-like constructs in line with the Humean theory of motivation. In this form, predictive

processing aligns more closely with other computational theories that maintain a belief-desire distinction, such as reinforcement learning and Bayesian decision theory.

The article also considers appeals to the free energy principle, which is sometimes used to argue for the elimination of desire-like constructs. It is argued that, on one reading of the free energy principle, it simply states that self-organizing systems can be redescribed *as if* they minimize free energy, and it thereby imposes no constraints on the actual mechanisms posited by process theories. On this interpretation, process theories are free to include desire-like constructs in their explanations of cognitive mechanisms.

The article concludes that predictive processing faces a dilemma: whether to prioritize parsimony of mental constructs or to offer a complete explanation of agency and the mind.

Author Contribution Statement

This dissertation consists of articles written and published during the course of my doctoral studies. Some of these articles were co-authored. Article 1, 'Is the wandering mind a planning mind?' was written in collaboration, Thor Grünbaum, while Article 4, 'Predictive minds can be Humean minds', was written in collaboration with Jelle Bruineberg and Thor Grünbaum. I was the first author and the primary contributor to each. I took the lead on identifying the research questions, formulating the central arguments, and structuring the overall narrative of the articles. In all cases, most of the core claims, theoretical framing, and argumentative strategies originated with me and reflect my independent line of inquiry. I was responsible for the drafting and redrafting of the texts and final editing. All submissions and revisions for peer review were carried out by me, in consultation with my co-authors where appropriate.

These collaborations were invaluable for sharpening the ideas and improving the clarity and rigor of the texts, but the articles themselves—and the broader research trajectory they collectively represent—are mostly the result of my independent scholarly work. Each piece contributes to the aims of the dissertation, which I conceived and pursued independently.

Parts of this dissertation have been published previously and appear here with minor revisions.

Article 1: Junker, F. T., & Grünbaum, T. (2024). Is the wandering mind a planning mind? *Mind & Language,* 39(5):706–725. doi: https://doi.org/10.1111/mila.12503

Article 4: Junker, F. T., Bruineberg, J., & Grünbaum, T. (2024) Predictive Minds Can Be Humean Minds. *The British Journal for the Philosophy of Science*.
doi: https://doi.org/10.1086/733413

References

Arpaly, N. & Schroeder, T. (2012). Deliberation and Acting for Reasons. *Philosophical Review, 121*:209-239.

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, *20*:1604–1611.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, *23*:1117–1122.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1):1-18.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Broome, J. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

Buckner, C. (2019). Rational Inference: The Lowest Bounds. *Philosophy and Phenomenological Research*, *98*(3):697-724.

Buehler, D. (2022). Agentive capacities: a capacity to guide. *Philosophical Studies, 179*(1):21-47.

Camp, E. (2007). Thinking with Maps. *Philosophical Perspectives, 21*:145–82

Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague, (eds.), *Non-Propositional Intentionality* (pp. 19-45). Oxford University Press.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford University Press.

Carruthers, P. (2025). *Explaining our Actions: A Critique of Common-Sense Theorizing*. Cambridge University Press.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive control system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106*:8719-8724.

Christoff, K., Irving, Z., Fox, K., Spreng, N., Andrews-Hanna, J. (2016). Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, *17*:718–731.

Christoff, K., Mills, C., Andrews-Hanna, J. R., Irving, Z. C., Thompson, E., Fox, K. C. R., & Kam, J. W. Y. (2018). Mind-Wandering as a Scientific Concept: Cutting through the Definitional Haze. *Trends in Cognitive Sciences*, *22*(11):957–959.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*:1-15.

Colombo, M. (2017). Social Motivation in Computational Neuroscience. In J. Kiverstein (ed.), *The Routledge Handbook of Philosophy of the Social Mind* (pp. 320–40). Routledge.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy, 60*(23):685.

Fodor, J. A. (1987). Why There Still Has to Be a Language of Thought. In *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (pp. 135–154). MIT Press.

Fodor, F. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1-2):3-71.

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage*, *111*:611–621.

Frankfurt, H. G. (1978) The Problem of Action. *American Philosophical Quarterly*, *15*(2): 157–162.

Gable, S. L., Hopper, E. A., & Schooler, J. W. (2019). When the muses strike: Creative ideas of physicists and writers routinely occur during mind wandering. *Psychological Science, 30*(3):396–404.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science, 317*(5843):1351–1354.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis and S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (pp. 623-653). MIT Press.

Haas, J. (2023). The evaluative mind. In J. Haugeland, C. F. Craver, & C. Klein (eds.), *Mind Design III: Philosophy, Psychology, and Artificial Intelligence* (pp. 295-313). MIT Press

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Irving, Z. C. (2016). Mind-wandering is unguided attention: accounting for the "purposeful" wanderer. *Philosophical Studies, 173*(2):547-571.

Irving, Z. C. (2021). Drifting and directed minds: The significance of mind-wandering for mental agency. *The Journal of Philosophy, 118*:614-644.

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: an experience-sampling study of working memory and executive control in daily life. *Psychological Science*, *18*(7):614–621.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*:932.

Klein, C. (2018). What do predictive coders want? *Synthese, 195*:2541-2557.

Klein, C. (2020). A Humean Challenge to Predictive Coding. In D. Mendonça, M. Curado & S.S. Gouveia (eds.), *The Philosophy and Science of Predictive Processing* (pp. 25–38). Bloomsbury Academic.

Levy, Y. (2024). Who is a reasoner? *Inquiry*, 1–27.

Mele, A. R. (2003). *Motivation and Agency*. Oxford University Press.

Munroe, W. (2021). Reasoning, rationality, and representation. *Synthese, 198*(9):8323-8345.

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences, 20*:605–617.

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018). Mind-Wandering as a Natural Kind: A Family-Resemblances View. *Trends in Cognitive Sciences*, *22*(6):479–490.

Murray, S. (2025). Vigilance and mind wandering. *Mind & Language, 40*(2):174-194.

Mylopoulos, M., & Pacherie, E. (2019). Intentions: The Dynamic Hierarchical Model Revisited. *Wiley Interdisciplinary Reviews. Cognitive Science, 10*(2):e1481.

Pacherie, E. (2008). The Phenomenology of Action: A Conceptual Framework. *Cognition*, *107*(1):179-217.

Pacherie, E., & Mylopoulos, M. (2020). Beyond Automaticity: The Psychological Complexity of Skill. *Topoi, 40*(3):649-662.

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature neuroscience*, *24*(9):1292–1301.

Parr, T., Pezzulo, G., & Friston, K. J., (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

Piñeros Glasscock, J. S. & S. Tenenbaum (2023). Action. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, URL <https://plato.stanford.edu/archives/spr2023/entries/action/>.

Quilty-Dunn, J. & Mandelbaum, E. (2018). Inferential Transitions. *Australasian Journal of Philosophy*, *96*(3):532-547.

Quilty-Dunn J., Porot, N., & Mandelbaum, E. (2023) The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences, 46, e261*: 1–75.

Railton, P. (2017). At the Core of Our Capacity to Act for a Reason: The Affective System and Evaluative Model-Based Learning and Control. *Emotion Review*, 9 (4):335-342.

Shea, N. (2024a). *Concepts at the Interface*. Oxford University Press.

Shea, N. (2024b). Metacognition of Inferential Transitions. *The Journal of Philosophy*, *121*:(11):597-627.

Siegel, S. (2019). Inference Without Reckoning. In M. B. Jackson & B. Jackson (eds.), Reasoning: New Essays on Theoretical and Practical Thinking (pp. 15-31). Oxford University Press.

Smith, M. (1987). The Humean theory of motivation. *Mind, 96*(381):36-61.

Turnbull, A., Wang, H. T., Murphy, C., Ho, N. S. P., Wang, X., Sormaz, M., Karapanagiotidis, T., Leech, R. M., Bernhardt, B., Margulies, D. S., Vatansever, D., Jefferies, E., & Smallwood, J. (2019). Left dorsolateral prefrontal cortex supports context-dependent prioritisation of off-task thought. *Nature Communications*, *10*(1).

Railton, P. (2017). At the Core of Our Capacity to Act for a Reason: The Affective System and Evaluative Model-Based Learning and Control. *Emotion Review*, *9*(4):335-342.

Rescorla, M. (2009a) Cognitive Maps and the Language of Thought, *British Journal for the Philosophy of Science, 60:*377–407.

Rescorla, M. (2009b). Predication and cartographic representation. *Synthese, 169*(1):175 - 200.

Rescorla, M. (2017). 'Maps in the Head?'. In Kristin Andrews & Jacob Beck (eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 34–45). Routledge.

Rescorla, M. (2023). The Language of Thought Hypothesis. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* , URL = <https://plato.stanford.edu/archives/sum2024/entries/language-thought/>

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*:487–518.

Shepherd, J. (2019). Why does the mind wander? *Neuroscience of Consciousness, 2019*.

Shepherd, J. (2025). Salvaging the "sense of agency": Metacognitive feelings for flexible behavioral control. *The Journal of Philosophy*.

Sripada, C. (2018). An exploration/exploitation tradeoff between mind wandering and goal-directed thinking. In K. C. Fox, & K. Christoff (eds.), *Oxford Handbook of Spontaneous Thought and Creativity*. Oxford University Press.

Sripada, C. (2025). The valuationist model of human agent architecture. *Philosophical Psychology*, 1–30.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica, 136*(3):370–381.

Stawarczyk, D., Cassol, H., & D'Argembeau, A. (2013). Phenomenology of future-oriented

mind-wandering episodes. *Frontiers in Psychology*, *4*, 425.

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5):1249–1263.

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). How to build a cognitive map. *Nature neuroscience*, *25*(10):1257–1272.

Wu, W. (2016). Experts and Deviants: The Story of Agentive Control. *Philosophy and Phenomenological Research, 93*(1):101-26.

ARTICLE 1

Is the Wandering Mind a Planning Mind?

Abstract: Recent studies on mind-wandering reveal its potential role in goal exploration and planning future actions. How to understand these exploratory functions and their impact on planning remains unclear. Given certain conceptions of intentions and beliefs, the exploratory functions of mind-wandering could lead to regular reconsideration of one's intentions. However, this would be in tension with the stability of intentions central to rational planning agency. We analyze the potential issue of excessive reconsideration caused by mind-wandering. Our response resolves this tension, presenting a model that aligns the roles of mind-wandering in planning with empirical evidence and the sustained stability of intentions.

1. Introduction

Recent empirical work on mind-wandering suggests that it might have various functional roles, including in autobiographical planning (Baird et al., 2011; Klinger, 2013; Stawarczyk et al., 2011, 2013) and creative problem-solving (Baird et al., 2012; Ruby et al., 2013; Fox & Beaty, 2019; Gable et al., 2019). This has led some to suggest that mind-wandering might be an exploratory process, allowing agents to explore new and potentially better opportunities (Sripada, 2018) or to search for more rewarding goals when the value of current goals is expected to be low (Shepherd, 2019). In this article, we review recent work on the functions of mind-wandering and develop a novel account of its role in planning. Our account will be motivated partly by philosophical theorizing and partly by empirical work.

Our starting point will be the suggestion that mind-wandering has an exploratory function. If mind-wandering often involves switching from exploiting existing goals to an exploratory mode of thought where new goals are assessed in the mind, it might involve some process in which current intentions are evaluated and possibly discarded. Add to this the observation that mind-wandering is ubiquitous. According to some estimates, we spend up to half of our waking hours mind-wandering (Killingsworth & Gilbert, 2010). This raises the possibility that reconsideration of one's intentions happens regularly. Yet this conflicts with a central assumption of the influential planning theory of intention (henceforth, PTI; Bratman, 1987), according to which intentions remain fairly stable over time and reconsideration should be rare.

According to PTI, future-oriented intentions are partial plans of action that play fundamental roles in deliberation and help coordinate our projects over time and with other agents. By committing ourselves to action in advance, we are able to make rational decisions in situations where we have too little time to deliberate, or it is too costly to do so. If this picture is correct, it helps explain how planning agents can make the best possible use of finite time and limited cognitive resources. But for this to work, the agent's prior intentions must remain relatively stable over time, that is, they must resist reconsideration. This we will refer to as the intention stability assumption. An agent that regularly reconsiders would likely tend to give up their intentions before the time to act on them arrives and so would have little to gain from committing themselves in advance compared to simply deliberating about what to do immediately before acting. Worse still, regular reconsideration risks undermining the benefits of committing and sticking to long-term projects and being reliable and predictable collaborators. To obtain these benefits, we trade off flexibility for stability. The worry is that exploratory mind-wandering might introduce too much flexibility.

The aim of the article is to discuss the proper characterization of the role of mind-wandering in planning. In Section 2, we introduce the hypothesis that mind-wandering has an exploratory function. In Section 3, we show that if this implies that mind-wandering leads to regular reconsideration, it is in tension with the PTI. In Section 4, we show that reconsideration is not the only planning-related function attributable to mind-wandering and in Section 5, we use this analysis to argue that mind-wandering does not lead to excessive reconsideration. In Section 6, we discuss how mind-wandering might alter our stock of beliefs and whether this makes it rational to regularly reconsider one's intentions. We argue that under certain models of rational formation and revision of intentions and beliefs, mind-wandering is unlikely to make regular reconsideration rational. Finally, in Section 7, we discuss the relationship between mind-wandering and active deliberation and explain how they are distinct, despite sharing certain functions such as attitude change.

2. Mind-Wandering as Mental Exploration

Some might find surprising the proposal that mind-wandering has a goal-directed dimension. The extent of one's surprise might depend on one's notion of mind-wandering. While we remain uncommitted to any particular conception and operationalization of mind-wandering in this article, we rely mainly on empirical studies that operationalize mind-wandering as task-unrelated or stimulus-independent thought. It remains to be seen exactly how these approaches

relate to alternative operationalizations, such as freely-moving thought (Mills et al., 2018) informed by the dynamic framework of thought (Christoff et al., 2016). Proponents of the dynamic framework (Christoff et al., 2016; Irving, 2016, 2021) sometimes emphasize the difference between mind-wandering, a relatively unconstrained and freely-moving mode of thought, and planning, a more constrained, deliberate, and goal-directed mode of thought. We suggest that the border between these modes of thought is not quite so categorical in that mind-wandering might make certain contributions to planning and deliberation without itself being highly constrained or deliberate.

Why should we think that mind-wandering plays a role in planning? Growing evidence suggests that rather than being a mere failure to control our thoughts, mind-wandering can be a strategy. Studies show that our thoughts frequently wander to information that is future-oriented, self-related, and goal-relevant thus potentially allowing us to anticipate personally relevant future goals (Baird et al., 2011; Stawarczyk et al., 2011, 2013). Mind-wandering can be swiftly and strategically modulated in anticipation of changes in task demands (Seli et al., 2018) and improve performance on social problem-solving (Ruby et al., 2013) and creative thinking tasks (Baird et al., 2012; Fox & Beaty, 2019; Gable et al., 2019). Hence, sometimes the best use of our cognitive resources (e.g., attention, working memory, and executive control) might be to let our minds wander.

Mind-wandering is also linked to episodic thought, that is, the ability to reconstruct events from one's personal past and to imagine counterfactual and possible future scenarios. Both mind-wandering and the various forms of episodic thought have self-generated content and activate the default mode network (Fox et al., 2015; Smallwood & Schooler, 2015). According to recent work, we flexibly recombine information from past experiences to construct simulations of what could have happened in the past or what may happen in the future (Schacter et al., 2007; De Brigard, 2014). Episodic simulation seems to be implicated in far-sighted decision-making, emotion regulation, prospective memory, and spatial navigation (Schacter et al., 2015, 2017). During mind-wandering, we also tend to generate episodic simulations (Baird et al., 2011; Stawarczyk et al., 2011, 2013). Other findings show that when coupled with the frontoparietal control network, the default mode network supports autobiographical planning, that is, the ability to identify and organize the steps needed to arrive at a certain personal future event (Spreng et al., 2010), and that mind-wandering shows similar coupling between the default mode network and executive areas (Fox et al., 2015). Together, these findings make it plausible that mind-wandering has mechanisms and functions in common with other kinds of episodic thought, including a role in planning.

But do the benefits of mind-wandering outweigh its costs? Studies show that mind-wandering can negatively affect performance on tasks that require monitoring and encoding of immediate input (e.g., comprehension during reading and lectures) and demanding tasks that require general intellectual functioning and executive control (e.g., sitting exams; Smallwood & Schooler, 2015). To reduce its costs, mind-wandering should be regulated in a context-dependent manner (Smallwood & Andrews-Hanna, 2013; Smallwood & Schooler, 2015), making it more common in non-demanding contexts and less common in tasks that require focused attention. This pattern is borne out by the evidence (Turnbull et al., 2019; Konu et al., 2021; Smallwood et al., 2021; Mulholland et al., 2023) with lower rates of mind-wandering during undemanding tasks in individuals with higher working memory capacity (Levinson et al., 2012). Moreover, the content of mind-wandering should tend to be future-oriented to allow agents to anticipate and plan for the future rather than past-directed which is associated with negative mood (Killingsworth & Gilbert, 2010; Smallwood & Schooler, 2015). These are exactly the patterns we find (Baird et al., 2011; Stawarczyk et al., 2011, 2013) with a stronger prospective bias in individuals with higher working memory capacity (Baird et al., 2011).

In sum, mind-wandering likely plays a significant and occasionally adaptive role in cognition, including in processes having to do with planning. One way to capture this is to think of mind-wandering as a kind of mental exploration. We might sometimes switch to mind-wandering to explore new options, because it is not always optimal to continue exploiting the same known options. That is, there might exist an exploration–exploitation tradeoff between mind-wandering and goal-directed thinking.

One such account comes from Joshua Shepherd (2019). Shepherd builds on the expected value of control theory of executive control according to which the executive control system determines how much control to exert toward specific goals based on a rational cost–benefit analysis (Shenhav et al., 2017). Specifically, the executive system tries to estimate which package of control signals (e.g., dictating what to attend to and how intensely) has the highest expected value of control, that is, strikes an optimal balance between expected gains (e.g., reward rate) and expected cost (including intrinsic costs to exerting control and opportunity costs of pursuing some strategies over others). According to Shepherd, the optimal package of control signals sometimes causes a switch to exploration, that is, a search for new and better goals, and sometimes mental exploration (e.g., querying memory) is deemed more cost-effective than exploring the environment.

Shepherd proposes that the function of mind-wandering might be such mental exploration: when the current goal is deemed insufficiently rewarding, the executive system initiates a

search for a new, more rewarding goal. Shepherd limits his discussion to unintentional mind-wandering, which he describes as 'episodes of mind-wandering that are neither initiated nor governed by any reportable intention of the agent' (p. 2) and posits that the agent is not conscious of the executive control mechanism directing the content of their stream of consciousness in a different direction. This is in line with research suggesting that mind-wandering is characterized by a lack of meta-awareness, that is, awareness of the current contents of one's stream of consciousness, including that one's mind is wandering (Smallwood, 2013; Smallwood & Schooler, 2015). Shepherd allows that mind-wandering might sometimes be completely unguided or guided in other ways (e.g., by affectively salient stimuli or other distractors) and sometimes happen consciously. Yet even when mind-wandering episodes are unguided, the executive system should, Shepherd suggests, be able to commandeer them for guided mental exploration when a valuable goal becomes salient.

A similar proposal has been made by Chandra Sripada (2018). According to Sripada, mind-wandering has the exploratory function of increasing informational stores and potentially opening up new opportunities for action. He proposes three possible accounts of the switching mechanism. First, mind-wandering might be the default state, which the mind switches to when goal-directed thinking is not required. Second, the brain might be wired to oscillate between wandering and goal-directed states at an appropriate rate to reap the benefits of each and avoid being stuck in either. And third, it might be that goal-directed thinking exhibits diminishing marginal utility over time, because after a certain amount of time, additional efforts are expected to be increasingly unlikely to yield additional gains. Thus, at some point, it becomes favorable to switch to mind-wandering to generate new information and creative insights. The agent might find it increasingly effortful to continue with goal-directed thinking and become increasingly prone to switch to mind-wandering which is experienced as less effortful. Like Shepherd, Sripada posits that the mechanisms leading to exploratory mind-wandering are unconsciously and unintentionally implemented.

3. Exploratory Mind-Wandering and Reconsideration

While recent research on mind-wandering and related mental phenomena supports the hypothesis that mind-wandering is involved in planning, it is not clear how it interfaces with philosophical work on planning. To close this gap, we propose an interpretation of exploratory mind-wandering from the perspective of rational planning agency. We begin by considering

the possible connection between exploratory mind-wandering and reconsideration of one's practical commitments and intentions.

According to Shepherd (2019), mind-wandering functions as a search for new and better goals. A goal can be understood as something the agent intends to achieve. One implication of this search might therefore be that the agent opens up the question about whether to do as previously intended: she comes to reconsider her prior intention. Sripada's (2018) account can be interpreted similarly. On his account, mind-wandering can increase informational stores to potentially open up new opportunities. Again, we might say this process could lead the agent to open up the question of whether to act as previously intended, that is, to reconsider prior intentions. If exploratory mind-wandering does indeed lead to reconsideration, this could have profound implications for rational planning agency. Specifically, if exploratory mind-wandering leads to regular reconsideration, this would conflict with the intention stability assumption of the PTI. We might state the problem as follows:

The Problem of Excessive Reconsideration

1. Mind-wandering has an exploratory function.
2. Exploratory mind-wandering can lead to reconsideration (assumes 1).
3. If exploratory mind-wandering happens regularly, then reconsideration happens regularly (assumes 2).
4. If reconsideration happens regularly, then the intention stability assumption of the PTI is false (from the definition of intention stability in the PTI).
5. Exploratory mind-wandering happens regularly (assumes 1).
6. Reconsideration happens regularly (from 3 and 5).
7. Conclusion: the intention stability assumption of the PTI is false (from 4 and 6).

Since many researchers have since built on the insights of the PTI, this would be disruptive for an entire research program. Thus, the argument also highlights the significance of mind-wandering research for research on planning agency. Yet the planning-related functions of mind-wandering might be more multifaceted than suggested by this argument.

4. The Many Faces of Exploratory Mind-Wandering

Thus far, we have only considered reconsideration, but there are other roles mind-wandering could play in planning. In this section, we provide a deeper analysis of the various planning-

related functions mind-wandering might serve. This will in turn allow us to formulate various responses to the problem of excessive reconsideration.

4.1. Three kinds of reconsideration

To properly assess if mind-wandering can lead to reconsideration—and if so, how regularly—we need to consider which forms of reconsideration mind-wandering can plausibly take. We will consider three different kinds: deliberative, policy-based, and non-reflective reconsideration (cf. Bratman, 1987, Chap. 5).

First, there is deliberative reconsideration. Here, the agent deliberates about whether to reconsider and decides to reconsider, which might result in either reaffirming her prior intention or canceling it. Might agents deliberate about whether to reconsider and decide to use mind-wandering as a means to do so? In that case, the agent would seem to engage in such mind-wandering intentionally. According to some studies, people report that they often do intentionally let their minds wander (Seli et al., 2016). However, there are reasons to think that mind-wandering cannot take the form of deliberative reconsideration. First, the coherence of intentional mind-wandering is itself controversial (Murray & Krasich, 2020). Second, existing accounts of intentional mind-wandering seem to rule out this type of deliberative reconsideration. According to Santiago Arango-Muñoz and Juan Pablo Bermúdez (2021), intentional mind-wandering is the intentional omission to control one's thoughts, specifically, 'the control required to string thoughts together toward the completion of a goal' (p. 7738). On Zachary Irving's (2021) account, intentional mind-wandering amounts to a type of meta-control where one monitors and regulates one's thinking to ensure that one's mind is wandering freely rather than fixating on a specific topic. However, during deliberative reconsideration one is in fact guiding one's thoughts toward the completion of an occurrent goal: to figure out whether to reaffirm or cancel one's intention and, in doing so, guiding one's thoughts toward content considered relevant to settling this question—thus fixating on a specific topic.

Second, there is policy-based reconsideration. This is when an agent adopts a policy to reconsider if certain conditions obtain. Perhaps agents can form a general policy to let their minds wander in certain situations (e.g., when their goals have proved unsuccessful or shown diminishing returns for some time) in the hopes of thinking of either reasons to reaffirm their current intention or of better alternatives and reasons for adopting them instead. However, this proposal is confronted with the same problem as the deliberative case. When the relevant circumstances arise and the agent begins to reconsider—as prescribed by the policy—the agent

(implicitly) adopts the goal of figuring out whether to reaffirm or cancel her current intention. The ensuing thought process of trying to achieve this goal will not be one of mind-wandering.

Finally, we have non-reflective reconsideration. This happens when an agent starts to seriously consider options incompatible with her prior intentions because of certain habits, skills, or dispositions (e.g., to notice certain problems or salient features of the environment) rather than through explicit deliberation. The agent thereby implicitly reopens the question of whether to do as previously intended. It seems plausible that we have a disposition to sometimes respond in this way to the propositions entertained during mind-wandering.

Consider an example: while mind-wandering, Esme thinks of a festival she would like to attend and that it takes place the same week that she plans to go hiking with her friend. Knowing that it might be possible to reschedule with her friend, she implicitly reopens the question of whether to go hiking that week. It might be argued that as Esme starts to weigh reasons for and against sticking to her original intention, she will be guiding her thoughts toward the completion of the goal of figuring out whether to reaffirm or cancel her original intention.

At this stage, she is no longer mind-wandering. However, by making salient a conflict between her various interests, her mind-wandering still provided the initial reason to reconsider and so leads her to reconsider, even if the subsequent weighing of reasons no longer counts as mind-wandering. The possibility of such cases suggests that we cannot rule out mind-wandering-induced reconsideration. To rule out excessive reconsideration, we therefore need to rule out that such reconsideration is excessive. One way to do this is to show that reconsideration rarely results from mind-wandering, since the information it generates typically supports rather than challenges our existing intentions.

While the example considered above focused on distal or ultimate goals of the agent, the evidence suggests that most mind-wandering episodes relate to goals that are more proximal. According to one study by Stawarczyk et al. (2013), 38% of future-oriented mind-wandering episodes relate to what will happen later in the present day and 27% to what will happen between tomorrow and the next 7 days. To the extent that mind-wandering leads to reconsideration, it should therefore be more prone to make us reconsider more proximal goals than distal ones. However, an alternative explanation is that mind-wandering is more likely to influence temporally closer sub-goals than make us reconsider distal goals.

4.2. Filling out of partial plans

Consistent with this last suggestion, an alternative construal of the exploratory function of mind-wandering is that mind-wandering helps fill out partial plans by exploring relevant means, preliminary steps, and more specific courses of action. This is supported by the evidence cited above suggesting that mind-wandering plays a role in autobiographical planning. Given looser constraints on its content, mind-wandering enables the consideration of a broader set of possibilities than more constrained, goal-directed thinking (Christoff et al., 2016; Irving, 2016, 2021). If this occasionally inspires better strategies, including better suggestions for how to fill out partial plans than would otherwise have been considered, this could explain the time and resources spent mind-wandering by a planning agent. According to the PTI, there is a rational requirement of means-end coherence such that when we intend a certain end and believe something to be a necessary means to achieve that end, we should also intend the means. This norm is pragmatically justified because abiding by it contributes to us getting what we (rationally) want in the long term (Bratman, 1987, Chap. 3). Thus, if mind-wandering makes us consider means to our ends, and we are rational planning agents, mind-wandering could bring us to intend such means.

Consider an example: Zara intends to go to the cinema with her friend this weekend, but they have not specified this plan further. As the weekend is only a few days away, her mind is prone to wander to this intention of hers and when it does, she starts thinking about which movie to see, which cinema to go to, and a few options spring to mind. Next, she starts thinking about calling her friend tonight to settle on a movie, place, and time and book the tickets before the good seats get taken. For the rest of the day, her mind tends to wander to these sub-goals, thus making it more likely that she will eventually become aware of the new options afforded to her during mind-wandering, consider them, commit to them, and ultimately execute them. We might add that these thoughts occurred to her during a moment of rest where she had no intention of thinking of anything in particular. Furthermore, the process was unconsciously implemented. As her mind wandered, she was unaware of it and did not intentionally guide her thoughts toward the completion of some particular goal (such as planning her weekend). This then seems like a paradigmatic example of mind-wandering. In addition, it also seems such mind-wandering helps the agent fill out a partial plan.

4.3. Reason-changing non-reconsideration

Another possibility is that mind-wandering might lead the agent to incorporate new considerations into her reasons for doing as she already intends without reconsidering those intentions. It seems plausible that this sometimes happens during mind-wandering. For example, during mind-wandering, Malik comes to think of an additional reason to visit his sister this week—something he has already decided to do—when he recalls that she is in the process of moving to a new flat and would no doubt appreciate his help. He does not reopen the question of whether to visit his sister (i.e., reconsider his intention) but his wandering mind changes his reasons for doing as he already intends.

Since neither filling out of partial plans nor finding new reasons for doing as one already intends entails reconsideration, they pose no threat to intention stability. Instead, they seem to support our commitment to and chances of successfully meeting our prior intentions. How does our account relate to that of Shepherd and Sripada? According to Shepherd (2019), exploratory mind-wandering consists in searching for new and better goals. Since filling out of partial plans might be understood as specifying sub-goals of more complex, distal goals, if we constrain the search to primarily specifying such sub-goals, Shepherd's account becomes compatible with ours. The discovery of new reasons for one's existing intentions is harder to construe as a search for goals, since reasons for action are (often) not themselves goals. On Sripada's (2018) account, mind-wandering increases informational stores to open up new opportunities for action. Nothing in this formulation seems to rule out that the new information and opportunities afforded by mind-wandering can support existing intentions by helping us fill out partial plans or discover new reasons for doing what we already intend.

5. Excessive Reconsideration Reconsidered

5.1. Does exploratory mind-wandering lead to (regular) reconsideration?

We are now in a position to respond to the problem of excessive reconsideration. One response would be to deny that mind-wandering has any exploratory function (against premise 1). However, the fact that mind-wandering tends to generate future-oriented, self-related, and goal-relevant information suggests that mind-wandering does allow us to explore new options that might lead to better outcomes in the long term. Another response would be to deny that mind-wandering ever leads to reconsideration (against premise 2). Yet the possibility of non-reflective reconsideration speaks against this. A more modest case can instead be made that

exploratory mind-wandering does not lead to regular reconsideration (against premise 3). We have argued that mind-wandering serves other planning-related functions. It might be that most exploratory mind-wandering serves to fill out partial plans or come up with new reasons supporting one's current intentions as opposed to triggering reconsideration.

Moreover, it might be argued that given the advantages to cognitively limited agents of forming plans ahead of time and sticking to them, it is implausible that mind-wandering would have evolved in a way that fundamentally undermined these advantages. More plausibly, the dispositions that might trigger reconsideration via mind-wandering are limited in scope (e.g., to infeasible, unimportant, or high-stakes intentions) so that they do not generally undermine the stability of our intentions. One might object that this begs the question, simply assuming that exploratory mind-wandering does not undermine rational planning agency on the grounds that being a rational planning agent is advantageous. So, what further reasons do we have for assuming that exploratory mind-wandering does not lead to regular reconsideration?

Here, we can appeal to two-tier accounts of rational (non)reconsideration (Bratman, 1987; Holton, 2009). Such accounts are designed to explain why it is rational for a planning agent not to reconsider in certain circumstances and avoid reconsideration in the face of prima facie triggers of reconsideration. The rationality of one's non-reconsideration (the lower tier) is assessed in terms of the rationality of the habit of non-reconsideration from which one's non-reconsideration follows (the higher tier). This is particularly important for explaining our tendency to resist non-reflective reconsideration for which there are many potential triggers, including thoughts we might have during mind-wandering. The two-tier approach states that an agent's non-reflective non-reconsideration of an intention is rational if it is the manifestation of general habits of non-reconsideration which are reasonable for the agent to have.

Michael Bratman (1987) argues that general habits of non-reconsideration explain our tendency not to reconsider our intentions in general. Having general habits of non-reconsideration is reasonable because it allows us to achieve complex projects that require long-term planning and vigilance and because it makes us more reliable partners when coordinating our plans with others which allows us to achieve more complex projects than we could individually. Richard Holton (2009) argues that the empirical literature bears out that we do in fact have such general habits of non-reconsideration and that such habits also provide the best explanation of our tendency not to reconsider our resolutions to resist temptation. However, occasional reconsideration is of course better than none. We should not be completely inflexible in light of changing and unexpected circumstances. We might have

corresponding habits of rational reconsideration which dispose us to reconsider when the stakes of our actions are high, or it is possible to deliberate in a low-cost, rational fashion.

Applied to exploratory mind-wandering, a case can now be made that we have a general presumption in favor of non-reconsideration even in the face of triggers of reconsideration, including those sometimes afforded by mind-wandering. However, when the stakes are sufficiently high or the opportunities afforded sufficiently great, we could be disposed to reconsider, and this would be rational under the circumstances. This view simultaneously allows that mind-wandering can occasionally lead to non-reflective reconsideration while remaining consistent with intention stability and rational planning agency. There is some evidence that mind-wandering supports non-reconsideration and intention stability. In one study, mind-wandering was associated with a greater capacity to resist the temptation of an immediate economic reward in favor of a larger future reward (Smallwood et al., 2013). According to a recent review, future-oriented mind-wandering tends to be about upcoming tasks and planned activities instead of novel hypothetical scenarios and mind-wandering about planned activities seems to increase the likelihood that these are accomplished (Kvavilashvili & Rummel, 2020). Thus, reflecting general habits of non-reconsideration, exploratory mind-wandering might be biased against reconsideration and toward filling out of partial plans and reason-changing non-reconsideration. But is a process biased against reconsideration in this way truly rational? We see no reason to deny this.

If such a bias allows cognitively limited agents to enjoy the dual fruits of mental exploration and rational planning agency, it might in fact be an optimal mental make-up for agents like us and thus no insult to rationality.

5.2. How regular is exploratory mind-wandering?

Finally, one could deny premise 5 of the problem of excessive reconsideration and claim that no empirical evidence supports the claim that exploratory mind-wandering is a common phenomenon. One might try to draw a distinction between exploratory and non-exploratory mind-wandering and argue that mind-wandering only rarely serves its exploratory function. It might be that the conditions necessary for mind-wandering to take the form of mental exploration only rarely obtain. What might such conditions be? First, we might say that mind-wandering is only exploratory when it is future-oriented because only future possibilities are relevant to our intentions and whether to reconsider them. Second, it might be argued that only mind-wandering with explicitly self-related and goal-relevant content serves its exploratory

function, since the purpose of exploration is to discover new information that the agent might exploit to improve her prospects.

There are, however, several problems with this argument. First, studies show that a quarter of mind-wandering episodes are reported as planning-related (Stawarczyk et al., 2013), future-oriented, self-related, and goal-relevant (Baird et al., 2011), suggesting that a lot of mind-wandering does bear on our intentions. Second, it is difficult to clearly delineate between stretches of mind-wandering that turn out useful and those that do not. We might not usually be aware of the potential utility of what we are experiencing during mind-wandering. For example, the new information might not seem immediately relevant to the agent but be stored in memory and become useful later. The agent can recall it during reasoning while remaining unaware that this information was first generated during mind-wandering. Admittedly, given the vagueness and uncertainty surrounding these distinctions and estimates, it is hard to precisely determine how often mind-wandering is genuinely exploratory.[4] But combined with the arguments above, we have good reason to doubt that exploratory mind-wandering leads to excessive reconsideration—even if we allow the occasional non-reflective reconsideration.

6. Changing Reasons

There is another way in which intention stability might come under threat from mind-wandering. We have suggested that mind-wandering might change the reasons the agent holds for doing as she intends without changing the intention itself. If these changes are significantly large, this eventually changes what intentions it is rational for the agent to hold. If the agent becomes aware of such changes to her reasons, she might realize that it is now rational for her to reconsider her intentions.

As suggested by Sripada (2018), a key function of mind-wandering might be to increase informational stores to potentially open up new opportunities. Mind-wandering might affect what information is available to the agent for processes like deliberation (among others), and, most relevant to our discussion, what beliefs the agent holds and is able to infer based on

[4] Recent studies using multidimensional experience sampling (Konu et al., 2021; Mulholland et al., 2023; Smallwood et al., 2021; Turnbull et al., 2019) have probed participants on multiple dimensions: temporal orientation, whether their thoughts were about themselves or others, whether they were thinking about solutions to problems (or goals), whether their thoughts were deliberate or spontaneous, whether they were thinking about one topic or many, whether their thoughts were about the environment or from memory, whether their thoughts were about something they already knew, and whether their thoughts were distracting from what they were doing, and other questions. This has enabled researchers to study which patterns of ongoing thought tend to arise in different task contexts, including during mind-wandering episodes. It would be interesting to see such methods brought to bear on whether mind-wandering makes the kinds of contributions to planning suggested here and, if so, in which task contexts and with what frequencies. One might add questions about whether the participants' thoughts led them to reconsider prior intentions, to fill out an existing plan, or to change their reasons for doing something they already intended to do. To our knowledge, no such study has been conducted. We would like to thank a reviewer for bringing these studies to our attention.

available information (e.g., about possible opportunities or goals). How does this relate to planning agency? For a planning agent to be rational she must only hold intentions that she believes it possible for her to execute (Holton, 2009, Chap. 3) or at least does not believe impossible to execute (Bratman, 1987, Chap. 3). Thus, were her beliefs to shift in such a way that now, according to those beliefs, it is either impossible or highly unlikely that she will be able to meet one of her intentions, it might now be rational to revise that intention.

Among the considerations relevant to whether we should revise an intention are relevant beliefs, such as whether we believe what we intend to do to be feasible or whether it might help us advance toward other ends we intend to achieve. We should therefore consider whether mind-wandering might change our beliefs to a point where, if we were to reflect on these changes, we should realize that the considerations supporting certain intentions have changed enough that we ought to reconsider those intentions to check if they are still supported by our reasons. As pointed out by Holton (2009, Chap. 1), a key feature of intention stability is that there are different thresholds for intention formation and revision. To ensure the stability of intentions, considerations sufficient to revise an intention must include significantly more relevant information than those sufficient to form it. The concern is therefore whether mind-wandering can surreptitiously generate a drift of beliefs large enough to regularly reach the threshold of rational reconsideration.

6.1. Doxastic effects of mind-wandering

But why should we believe that mind-wandering affects our beliefs? As mentioned above, mind-wandering often involves episodic simulation, which can affect beliefs in multiple ways. First, during episodic simulation, an agent may fill in gaps in memory with imagined or fictional details which might distort beliefs about past events (De Brigard, 2014). Second, counterfactual simulations of events that did not happen but could have might affect the agent's beliefs about causal relationships and probabilities. For example, simulating alternative causes or outcomes might lead to updated beliefs about what caused a particular event or what is likely to happen in similar future events. Third, episodic simulations can evoke emotional experiences and change the agent's beliefs about the desirability and plausibility of such events. One study shows that repeated simulation increases the perceived plausibility for emotional (positive or negative) future interpersonal experiences, but not neutral ones (Szpunar & Schacter, 2013). Another study indicates that repeated simulation of episodic counterfactual events decreases their perceived plausibility regardless of valence (De Brigard et al., 2013).

Thus, under the assumption that episodic simulations generated during mind-wandering have similar effects on beliefs, we have some inductive reasons to accept that mind-wandering affects beliefs. And since our beliefs about, say, what is likely to happen in the future or what the consequences of our actions might be partially constitute what we have reason to do, significant changes to such beliefs can change what intentions it is rational for us to hold (onto). For example, if someone intending to leave home without an umbrella gradually finds it more and more plausible that it will rain (perhaps through repeated simulations of the poor weather the past weeks), it eventually becomes rational for that person to reconsider whether to bring an umbrella. Large regular changes to the beliefs that guide our actions could make it rational for us to regularly reconsider, thus threatening intention stability. So, a key question is: Does mind-wandering cause large doxastic changes? Moreover, to determine whether mind-wandering supports or interferes with planning, we also need to consider whether the beliefs formed because of mind-wandering reliably help the agent meet her long-term goals.

6.2. Is thinking believing?

Recent discussions of belief acquisition, revision, and storage provide a good starting point for investigating these questions. Some theorists distinguish between Cartesian and Spinozan models of belief acquisition (Gilbert, 1991; Egan, 2008; Mandelbaum, 2014). On the *Cartesian model*, when we encounter a proposition (e.g., through the deliverances of perception or imagination), we can entertain a proposition without believing it and only assent to it (thus coming to believe it) after subjecting it to an evaluation that determines whether it should be accepted or rejected. By contrast, on the *Spinozan model*, we directly and automatically come to believe the propositions we process and only after subsequent effortful evaluation might we come to reject it.

Several conclusions have been drawn from the Spinozan view. First, the Spinozan model implies, and is meant to explain, that we harbor inconsistent beliefs, since on this model new beliefs are continuously acquired without evaluating whether they are consistent with our current stock of beliefs (Egan, 2008; Mandelbaum, 2014). Proponents of the Spinozan view have argued that this is best explained by a *fragmentation model* of beliefs according to which beliefs are stored in distinct, independently accessible fragments which are typically activated (to guide reasoning and action) and updated one at a time. Which fragment is activated—and so, in which fragment new beliefs are stored—depends on our current context. A fragmented belief system allows us to store inconsistent beliefs across different fragments even if the

beliefs stored within each fragment are kept consistent (Egan, 2008; Bendaña & Mandelbaum, 2021). This contrasts with a *unified model* of beliefs according to which beliefs are stored in a single database, reasoning and action is synchronically guided by the entire belief system, and belief revisions are sensitive to global properties of one's belief system such that when one belief changes, all other beliefs are (ideally) revised to remain consistent with the change.

Second, some proponents of Spinozan and fragmentation models argue that these best explain various ways in which our beliefs are biased with some perilous implications for rationality. According to Eric Mandelbaum (2014), the Spinozan view helps explain confirmation bias (i.e., our tendency to search for evidence that confirms our existing beliefs and resist evidence that disconfirms them). One puzzle about confirmation bias is that we sometimes experience cognitive dissonance even when we merely consider a proposition. If we automatically believe every proposition we consider, mere consideration will sometimes lead us to acquire beliefs that conflict with other standing beliefs, resulting in a dissonant state. The dissonant state is experienced as discomfort which reinforces dispositions to avoid searching for or calling to mind disconfirmatory evidence. Mandelbaum (2014) argues that this makes impartial deliberation impossible: whenever we consider a proposition, we come to believe it, thus making it susceptible to confirmation bias.

In addition, Mandelbaum (2019) argues that a core feature of belief revision is that it protects our self-image even at the expense of not being responsive to the evidence and not updating beliefs in a Bayesian way. When evidence contradicts subjectively important beliefs that constitute our self-image (e.g., that we are good, smart, and competent people), belief revisions resolve the resulting discomfort by protecting the subjectively important beliefs and resisting the conflicting evidence. According to Joseph Bendaña and Eric Mandelbaum (2021), this contrasts with a central assumption of unified models, namely, that the beliefs least open to revision are those whose revision requires the highest number of changes to other beliefs to keep the total belief system consistent (e.g., rules of logic or mathematics). Revising one's self-image, however, generally does require that one revises much else that one believes. Since fragmentation models are not committed to consistency across fragments, they can better accommodate such biased belief revision.

While none of these theories might be entirely true (e.g., maybe some belief-forming mechanisms are more Cartesian and some more Spinozan), they help capture general positions one can take on the nature of belief acquisition, revision, and storage and thus provide a useful starting point for theorizing about the doxastic effects of various mental phenomena. We will use these theories to make two points. First, the combination of Spinozan and fragmentation

models defended by some (Egan, 2008; Bendaña & Mandelbaum, 2021) is in tension with rational planning agency. Second, when applied to mind-wandering as a belief-forming mechanism, the combination of Spinozan and fragmentation models has even more troubling implications for rational planning agency.

Spinozan fragmentation models are in tension with rational planning agency in several ways. It is hard to see how the kind of rational deliberation conducive to successfully meeting our long-term goals is possible under this picture. According to the PTI, to accrue the benefits of long-term planning, we are rationally required to keep our intentions consistent with each other and with our beliefs. However, if we have a fragmented belief system containing many inconsistent beliefs, for many intentions there is likely to be some fragments with which the intention is consistent and some with which it is inconsistent. This is worrisome enough as it stands. But, if we accept that mind-wandering can lead us to acquire and revise beliefs, the threat to intention stability and rationality is exacerbated. Should we accept the antecedent? On the Spinozan view, as it is often stated, it seems that we must, since it does not discriminate between belief-forming mechanisms. Some proponents mention that it does not matter whether the proposition appears in perception or imagination (Gilbert, 1991; Mandelbaum, 2014).[5] Since we imagine many different propositional contents during mind-wandering, these staunch Spinozans should accept that mind-wandering can form new beliefs.

Due to the ubiquity of mind-wandering, this seems to entail a fast build-up of inconsistent beliefs. New and old fragments would continuously be opened, with new beliefs added or old ones revised, as more propositions are entertained by our wandering thoughts. If at one moment an agent's mind wanders to her resolution to stick to her diet and the next moment to worry that she will be tempted to order too much junk food, does that suffice to make her believe that she will do both things? Worse still, does this make her resolution a victim to which fragment happens to be active around dinnertime? The Spinozan might respond by limiting their view to certain modalities (e.g., perception) and accept that beliefs are not automatically acquired about propositions entertained during mind-wandering (and perhaps imagination more generally). On the other hand, if the Spinozan fragmentationist doubles down and accepts that mind-wandering can indeed open and reopen new and old fragments, this could lead to substantial drifts in the agent's belief sets, or fragments, over time.

Different lines of work support that mind-wandering could lead to large doxastic drifts given the truth of the Spinozan story. First, according to the decoupling hypothesis

[5] Egan (2008) restricts his discussion to perception.

(Smallwood, 2013; Turnbull et al., 2019), during mind-wandering, executive control processes disengage attentional processes from external stimuli, which insulates the internal stream of thought from perceptual distractions and ensures efficient processing of self-generated information. Second, work on the dynamics of mind-wandering suggests that mind-wandering episodes can vary widely in content and is characterized by a repeating pattern of a cluster of related thoughts about one topic followed by a jump to a new topic only modestly related to the previous one (Sripada & Taxali, 2020). Mind-wandering thus seems able to generate a diverse set of propositions linked to a wide set of contexts and so, if the Spinozan story is correct, to form and revise a large number of different beliefs. If large enough, such gradual drifts in the agent's belief sets might mean that the threshold for rational reconsideration is regularly crossed, thus making it irrational for the agent to avoid reconsideration of their intentions for very long.

One way even the staunch Spinozan could protect intention stability would be to argue that the threshold for rational reconsideration is very high indeed. But this seems equivalent to saying that the agent is highly insensitive to the fact that her beliefs might no longer support her intentions, which seems irrational. The more promising solution might be to argue that beliefs are more stable than the Spinozan would have it. Indeed, some have argued that planning benefits from keeping our beliefs reasonably stable so that we can reason and plan on the basis of them and remain committed to pursuing (difficult) long-term goals even in the face of a constant flux of new relevant evidence or setbacks. On one model, we are disposed to ignore some new evidence so as not to regularly reconsider our beliefs unless it passes a threshold beyond which it cannot properly be ignored (Holton, 2014). On another, we remain open to evidence that success on difficult long-term goals is not forthcoming but only reduces our confidence that continued effort will yield success when a certain evidential threshold has been passed (Morton & Paul, 2019). How high this threshold should be depends on the context and the agent's ability to bear the costs of failure.

Another concern is that some Spinozan fragmentationists (Mandelbaum, 2014, 2019; Bendaña & Mandelbaum, 2021) might overstate the extent of biased belief revision in a way that could be detrimental to effective long-term planning. If we constantly and automatically acquire new beliefs, and these were systematically prone to confirm our existing beliefs and protect our self-image, we would risk being left with highly partial and unreliable information about the prospects of success in our long-term goals. If, when our minds wander to how we might meet our long-term goals, we were prone to exaggerate our own competence and generally come to believe that success is forthcoming even when it is not, we would be unable

to properly assess when it is rational to stick to our guns and when it is rational to quit. A highly unreliable exploratory system risks being maladaptive, and it is unclear why mind-wandering would have evolved that way.

While we agree that human agents sometimes do exhibit the kinds of irrational behaviors that have motivated Spinozan fragmentation models, the proposed cure could be worse than the disease. Specifically, these models are in tension with some rational planning behaviors that we also seem to exhibit (even if fallibly so). However, it is important to note that not all fragmentationists are as pessimistic about our capacity for rationality as, for example, Bendaña and Mandelbaum (2021). Seth Yalcin (2021) tries to show that fragmentation per se is not irrational. Andy Egan (2008) argues that fragmentation might help guard against unreliable beliefforming mechanisms, since beliefs from unaffected fragments can help us infer that the outputs of certain mechanisms are unreliable. Cristina Borgoni (2021) suggests that even if we only keep beliefs consistent within fragments, we might still be responsive to evidence across fragments by having beliefs from inactive fragments stand as evidence for active fragments. Adam Elga and Agustín Rayo (2022) develop a version of fragmentation that is compatible with Bayesian decision theory. Still, a tension remains between having a fragmented belief system containing inconsistent beliefs and the rational requirement on planning agents to keep their intentions and beliefs consistent.

Since it is primarily the Spinozan view that entails large inconsistencies in a fragmented belief system, we have reason to doubt that beliefs are always acquired as automatically as the staunch Spinozan suggests. While we might be more prone to automatically believe what we perceive (Gilbert, 1991; Egan, 2008), we might be less prone to automatically believe what we imagine (e.g., during mind-wandering). Importantly, not all fragmentationists explicitly endorse the Spinozan view potentially leaving room for fragmentation with less inconsistency and less irrationality (e.g., Borgoni, 2021; Yalcin, 2021; Elga & Rayo, 2022).

We conclude that on pain of undermining long-term planning, agents are under rational pressure to reduce inconsistencies. Accepting that mind-wandering can change our beliefs reinforces this need. Furthermore, for mind-wandering to support effective planning, the beliefs we form about our prospects for success during mind-wandering should be at least somewhat reliable. While we hope to have raised some interesting epistemological questions about mind-wandering (e.g., about its reliability and whether updating beliefs based on self-generated information is justified), providing satisfactory answers to these will be a project for another time. Suffice it to say that it might be possible to form justified beliefs based on mind-wandering (e.g., if it turns out to be sufficiently reliable). If during mind-wandering, one is

reminded of multiple failed attempts at pursuing a similar goal in the past and that one's skills and odds of success have not improved since, it seems that one is justified in forming the belief that success this time around is unlikely. Yet a question remains about how exactly mind-wandering might change our beliefs. On the face of it, the Spinozan model seemed well-positioned to explain how mind-wandering can lead to the acquisition of beliefs, since it does not require the kind of reflective evaluation of the evidence that seems to be absent during mind-wandering. But due to its apparent tension with rational planning agency, we have reason to be skeptical of such a model—at least in the domain of mind-wandering and imagination. In the next section, we discuss how mind-wandering might lead to attitude change.

7. Mind-Wandering and Deliberation

So far, we have explained how exploratory mind-wandering might modify our intentions and beliefs and argued that despite the ubiquity of mind-wandering, this need not conflict with us being rational planning agents. Exploratory mind-wandering both contributes new considerations in support of our existing intentions and allows us to adapt to changing circumstances by updating our reasons for action. Given the multifaceted nature of mind-wandering, it should not be too surprising that it can serve such different functions. However, in trying to reconcile exploratory mind-wandering with rational planning agency, we seem to encounter another puzzle. The functions we have attributed to mind-wandering overlap with those standardly attributed to deliberation. But while deliberation seems to be constituted by a variety of different mental actions (shifting attention, inhibiting urges, imagining possible actions or outcomes, comparing options, weighing reasons, etc.), mind-wandering appears does not appear to be actively controlled in the same way.

We can resolve this apparent tension by explaining how mind-wandering relates to, yet remains distinct from, deliberation. Each of the planning-related functions we have attributed to mind-wandering—non-reflective reconsideration, completion of partial plans, and reason-changing non-reconsideration—can occur without some of the agentive features distinctive of deliberation, particularly intentional action. During mind-wandering, the execution of its planning-related functions is not something the agent is intentionally trying to bring about (and thus is not an intentional mental action, cf. Mele, 2009) nor is the agent intentionally directing their attention to particular pieces of information relevant to making a specific plan or decision (and thus it is not an intentional mental act of deciding, cf. Shepherd, 2015). Mind-wandering need only be guided in a way that enables the execution of its planning-related functions. To

achieve this, the agent need not intend to think of anything in particular and guide their thoughts toward the execution of any particular goal while correcting any deviation from this goal (for discussion, see Irving, 2016, 2021). Moreover, the guidance involved in mind-wandering allows that the same episode of mind-wandering can include thoughts related to various different goals as well as goal-irrelevant thoughts which, again, distinguishes it from intentional, deliberative thinking where we try to focus on one goal for an extended period of time and bring our attention back to the task when its strays to goal-irrelevant thoughts.

There are several ways in which mind-wandering might interact with deliberation while remaining distinct from it. This depends on the view one takes on the role of deliberation in action, specifically, whether action always requires deliberation. One might adopt the view that intentional action requires that one has previously deliberated about whether to perform the action in question, decided to perform the action, and thus intentionally formed an intention to perform the action. If we also assume that mind-wandering never itself constitutes deliberation (say, because of lacking certain agentive features), this has important implications for how to cash out the planning-related functions of mind-wandering. On a strict version of this view, mind-wandering cannot directly change our intentions or beliefs without intermediate deliberation. Instead, mind-wandering might at best trigger acts of deliberation that evaluate the potentially goal-relevant information generated during mind-wandering or encode new information that can be recalled during later acts of deliberation. For the information to change our attitudes and cause action, it might be argued, requires that it first be critically evaluated and integrated with other information during acts of deliberation. In other words, on a view where deliberation is necessary for attitude change and action, mind-wandering can only indirectly affect our attitudes and actions by generating inputs to deliberation.

However, there is reason to reject such a view. Some have argued that to avoid an infinite regress, there must be processes that allow us to think and act for reasons without deliberation. Since deliberation is an intentional mental action, if all actions required prior deliberation, all acts of deliberation would themselves require prior acts of deliberation ad infinitum. According to Nomy Arpaly and Timothy Schroeder (2012), non-deliberative, non-voluntary processes can still be reasons-responsive if mental transitions occur because certain logical relations (theoretical entailment, practical entailment, statistical relevance, etc.) obtain between the implicated attitudes. Such processes can involve transitions ‘from some beliefs to others (when believing for reasons), from beliefs (and perhaps desires and plans) to an intention or willed action (when acting for reasons), and perhaps other transitions as well’ (Ibid., p. 236). Exploratory mind-wandering is one candidate for such non-deliberative, non-voluntary yet

reasons-responsive processes. Thus, on this type of view, mind-wandering can bypass deliberation and directly change attitudes (e.g., form beliefs and intentions), which in turn changes what actions we are likely to perform.

Yet given the foundational role ascribed to non-deliberative, non-voluntary processes in Arpaly and Schroeder's account, this view risks ascribing too much importance to such processes relative to deliberation. On their account, deliberation plays the modest role of occasionally removing barriers to the non-deliberative, non-voluntary processes which are the real foundation for our ability to think and act for reasons. Deliberation, they argue, might for example refocus attention to deal with distraction, call to mind relevant information to deal with lack of inspiration, promote neglected facts that have not recently come to conscious attention, or sequence the stages of a difficult problem.

Even if we accept that mind-wandering can change attitudes in ways that are non-deliberative yet reasons-responsive, we need not accept that non-deliberative, non-voluntary processes are foundational in Arpaly and Schroeder's sense and that deliberation merely serves to remove barriers. There is room for an intermediate view on which deliberation is allowed a more substantial and independent role. For example, even if we accept that some intentions are acquired unintentionally, we have reason to believe that others are intentionally formed through acts of deciding, specifically, when we are uncertain or unsettled about what to do—and if the intentions to decide are themselves acquired unintentionally, there is no regress (Mele, 2003, Chap. 9). Nothing we have said rules out that various acts of deliberation still play a substantial role in, say, forming intentions in the face of uncertainty, explicitly and critically evaluating reasons, or changing attitudes in accordance with rational norms. In exploratory mind-wandering, we have identified a non-deliberative process that is poised to change our attitudes and how we act—possibly sometimes even in a reasons-responsive, rational way—thus further vindicating the existence and significance of such processes. Yet this does not replace deliberation so much as supplement it.

Since we can distinguish mind-wandering from deliberative processes, this account also seems broadly compatible with the distinction between unconstrained and constrained modes of thought proposed by the dynamic framework of thought (Christoff et al., 2016). An additional element in our account is that these modes interact as seen by the contributions mind-wandering makes to planning and deliberation.

How does the suggestion that mind-wandering is a non-deliberative process square with neural evidence that mind-wandering is regulated by executive areas (Turnbull et al., 2019)

known to be involved in deliberative processes (Botvinick & An, 2009)?[6] Regulation by executive areas is not, by itself, sufficient to render mind-wandering an act of deliberation. Deliberation is typically initiated with the intention of resolving an open question, and one's thoughts are guided toward that goal until it is settled. As noted in Section 4.1, such a thought process is not a form of mind-wandering, even on accounts of intentional mind-wandering. Thus, the kind of executive regulation involved in mind-wandering does not make it intentional or goal-directed in the manner required for deliberation, and the evidence remains compatible with the view that mind-wandering is non-deliberative.

8. Conclusion

We have argued that mind-wandering-based reconsideration should be a rare occurrence. Rather than prompting and rationalizing reconsideration, mind-wandering is more likely to help us fill out partial plans or think of new reasons for doing as we already intend. If this is the case, mind-wandering is unlikely to threaten intention stability. Another possibility is that mind-wandering could lead to gradual drifts in our beliefs over time that makes it rational to regularly reconsider our intentions. However, given reasonable thresholds for rational reconsideration and rationality-friendly models of belief acquisition, updating, and storage, mind-wandering is unlikely to induce drifts in our belief sets to an extent that makes regular reconsideration rational. Finally, we have tried to clarify the relationship between mind-wandering and active deliberation and shown that while the two serve similar functions and might interact, they remain distinct processes.

References

Arango-Muñoz, S., & Bermúdez, J. P. (2021). Intentional mind-wandering as intentional omission: The surrealist method. *Synthese, 199*:7727–7748.

Arpaly, N., & Schroeder, T. (2012). Deliberation and acting for reasons. *Philosophical Review, 121*:209–239.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W.

[6] Specifically, Turnbull et al. (2019) have shown that the dorsolateral prefrontal cortex (dlPFC) is involved in regulating mind-wandering in a context-dependent manner. Their proposal is that the dlPFC prioritizes task-relevant information by monitoring signals from internal and external sources and when external task-demands are high, the dlPFC suppresses mind-wandering. When demands are low, the dlPFC prioritizes mind-wandering by reducing the processing of external task-relevant signals and decoupling attention from external signals in order to facilitate efficient processing of self-generated information. We would like to thank a reviewer for bringing this to our attention.

(2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science, 23*:1117–1122.

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition, 20*:1604–1611.

Bendaña, J., & Mandelbaum, E. (2021). The fragmentation of belief. In C. Borgoni, D. Kindermann, & A. Onofri (eds.), *The Fragmented Mind* (pp. 78–107). Oxford University Press.

Borgoni, C. (2021). Rationality in fragmented belief systems. In C. Borgoni, D. Kindermann, & A. Onofri (eds.), *The Fragmented Mind* (pp. 137–155). Oxford University Press.

Botvinick, M., & An, J. (2009). Goal-directed decision making in prefrontal cortex: A computational framework. *Advances in Neural Information Processing Systems, 21*:169–176.

Bratman, M. (1987). Intention, Plans, and Practical Reason. Harvard University Press.

Christoff, K., Irving, Z., Fox, K., Spreng, N., & Andrews-Hanna, J. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*:718–731.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese, 191*:1–31.

De Brigard, F., Szpunar, K. K., & Schacter, D. L. (2013). Coming to grips with the past: Effect of repeated simulation on the perceived plausibility of episodic counterfactual thoughts. *Psychological Science, 24*:1329–1334.

Egan, A. (2008). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies, 140*:47–63.

Elga, A., & Rayo, A. (2022). Fragmentation and logical omniscience. *Noûs, 56*:716–741.

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage, 111*:611–621.

Fox, K. C. R., & Beaty, R. E. (2019). Mind-wandering as creative thinking: Neural, psychological, and theoretical considerations. *Current Opinion in Behavioral Sciences, 27*:123–130.

Gable, S. L., Hopper, E. A., & Schooler, J. W. (2019). When the muses strike: Creative ideas of physicists and writers routinely occur during mind wandering. *Psychological Science, 30*(3):396–404.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*:107–119.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press.

Holton, R. (2014). Intention as a model for belief. In M. Vargas & G. Yaffe (eds.), *Rational and Social Agency: The Philosophy of Michael Bratman* (pp. 12–37). Oxford University Press.

Irving, Z. C. (2016). Mind-wandering is unguided attention: Accounting for the "purposeful" wanderer. *Philosophical Studies, 173*:547–571.

Irving, Z. C. (2021). Drifting and directed minds: The significance of mind-wandering for mental agency. *The Journal of Philosophy, 118*:614–644.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science, 330*:932.

Klinger, E. (2013). Goal commitments and the content of thoughts and dreams: Basic principles. *Frontiers in Psychology, 4*:415.

Konu, D., Mckeown, B., Turnbull, A., Siu Ping Ho, N., Karapanagiotidis, T., Vanderwal, T., McCall, C., Tipper, S. P., Jefferies, E., & Smallwood, J. (2021). Exploring patterns of ongoing thought under naturalistic and conventional task-based conditions. *Consciousness and Cognition, 93*:103139.

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: A review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology, 24*(3):210–237.

Levinson, D. B., Smallwood, J., & Davidson, R. J. (2012). The persistence of thought: Evidence for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science, 23*(4):375–380.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy, 57*:55–96.

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language, 34*:141–157.

Mele, A. R. (2003). *Motivation and Agency*. Oxford University Press.

Mele, A. R. (2009). Mental action: A case study. In L. O'Brien & M. Soteriou (eds.), *Mental Actions* (pp. 17–37). Oxford University Press.

Mills, C., Raffaelli, Q., Irving, Z. C., Stan, D., & Christoff, K. (2018). Is an off-task mind a freely-moving mind? Examining the relationship between different dimensions of thought. *Consciousness and Cognition, 58*:20–33.

Morton, J. M., & Paul, S. K. (2019). Grit. *Ethics, 129*:175–203.

Mulholland, B., Goodall-Halliwell, I., Wallace, R., Chitiz, L., Mckeown, B., Rastan, A., Poerio, G. L., Leech, R., Turnbull, A., Klein, A., Milham, M., Wammes, J. D., Jefferies, E., & Smallwood, J. (2023). Patterns of ongoing thought in the real world. *Consciousness and Cognition, 114*:103530.

Murray, S., & Krasich, K. (2020). Can the mind wander intentionally? *Mind & Language, 37*:432–443.

Ruby, F. J., Smallwood, J., Sackur, J., & Singer, T. (2013). Is self-generated thought a means of social problem solving? *Frontiers in Psychology, 4*:962.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience, 8*:657–661.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory, 117*:14–21.

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences, 17*:41–50.

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences, 20*:605–617.

Seli, P., Carriere, J. S. A., Wammes, J. D., Risko, E. F., Schacter, D. L., & Smilek, D. (2018). On the clock: Evidence for rapid and strategic modulation of mind wandering. *Psychological Science, 29*:1247–1256.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience, 40*:99–124.

Shepherd, J. (2015). Deciding as intentional action: Control over decisions. *Australasian Journal of Philosophy, 93*:335–351.

Shepherd, J. (2019). Why does the mind wander? *Neuroscience of Consciousness, 2019.*

Smallwood, J. (2013). Distinguishing how from why the mind wanders: A process-occurrence framework for self-generated mental activity. *Psychological Bulletin, 139(3)*:519–535.

Smallwood, J., & Andrews-Hanna, J. (2013). Not all minds that wander are lost: The importance of a balanced perspective on the mind-wandering state. *Frontiers in Psychology, 4*:441.

Smallwood, J., Ruby, F. J., & Singer, T. (2013). Letting go of the present: Mind-wandering is associated with reduced delay discounting. *Consciousness and Cognition, 22(1)*:1–7.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically

navigating the stream of consciousness. *Annual Review of Psychology, 66*:487–518.

Smallwood, J., Turnbull, A., Wang, H. T., Ho, N. S. P., Poerio, G. L., Karapanagiotidis, T., Konu, D., Mckeown, B., Zhang, M., Murphy, C., Vatansever, D., Bzdok, D., Konishi, M., Leech, R., Seli, P., Schooler, J. W., Bernhardt, B., Margulies, D. S., & Jefferies, E. (2021). The neural correlates of ongoing conscious thought. *iScience, 24(3)*:102132.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage, 53*:303–317.

Sripada, C. (2018). An exploration/exploitation tradeoff between mind wandering and goal-directed thinking. In K. C. Fox, & K. Christoff (eds.), *Oxford Handbook of Spontaneous Thought and Creativity*. Oxford University Press.

Sripada, C., & Taxali, A. (2020). Structure in the stream of consciousness: Evidence from a verbalized thought protocol and automated text analytic methods. *Consciousness and Cognition, 85*:103007.

Stawarczyk, D., Cassol, H., & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology, 4*:425.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica, 136(3)*:370–381.

Szpunar, K. K., & Schacter, D. L. (2013). Get real: Effects of repeated simulation and emotion on the perceived plausibility of future experiences. *Journal of Experimental Psychology. General, 142*:323–327.

Turnbull, A., Wang, H. T., Murphy, C., Ho, N. S. P., Wang, X., Sormaz, M., Karapanagiotidis, T., Leech, R. M., Bernhardt, B., Margulies, D. S., Vatansever, D., Jefferies, E., & Smallwood, J. (2019). Left dorsolateral prefrontal cortex supports context-dependent prioritisation of off-task thought. *Nature Communications, 10(1)*:3816.

Yalcin, S. (2021). Fragmented but rational. In C. Borgoni, D. Kindermann, & A. Onofri (eds.), *The Fragmented Mind* (pp. 156–180). Oxford University Press.

ARTICLE 2

Mind-Wandering in Action

Abstract: Mind-wandering is often considered passive, yet it also plays a role in advancing our goals. Some have sought to reconcile this tension by suggesting that mind-wandering is at once passive and purposive. This compromise overlooks that explaining how mind-wandering contributes to goal pursuit requires rejecting its passivity. What explains mind-wandering's contribution to goal pursuit are mechanisms that actively monitor, evaluate, and regulate mental representations and processes. In this light, mind-wandering emerges as an actively guided learning process that informs conclusions about what is the case or what to do. This carries significant implications for our understanding of rational inference and mental agency.

1. Introduction

Imagine yourself strolling through town, not trying to think about anything in particular. Still, your mind is anything but quiet. In rapid succession, your thoughts drift from a tricky problem at work to a tasteless joke you overheard, to a stranger's kind gesture, and to what you might cook for dinner. This familiar mode of thought—commonly referred to as mind-wandering—is ubiquitous, occupying 30-50% of our waking hours (Kane et al., 2007; Killingsworth & Gilbert, 2010). It appears, at once, passive and unconstrained, yet is often helpful in advancing our goals and has attracted growing interest among philosophers who have tried to reconcile its passive and purposive nature. In this paper, I argue that this is a mistake. The standard view that mind-wandering is both passive and purposive masks a tension: to uphold one, we must revise or abandon the other. A satisfactory explanation of mind-wandering's role in goal pursuit compels us to relinquish the view that it is passive.

A commonly touted desideratum is that we must explain the passive nature of mind-wandering, often expressed along the following lines:

> *The Passivity Thesis*: mind-wandering is something that merely happens to us, as opposed to something we actively do.

To motivate this idea, one might appeal to the commonsensical notion that we experience ourselves as passive receivers of wandering thoughts rather than as agents in control of them.

One might also point to the absence of relevant marks of agency, as articulated by philosophical accounts of agency, or empirical differences between mind-wandering and paradigmatically active, goal-directed forms of thinking. In particular, mind-wandering's meandering character—the tendency to frequently jump from one topic to another, often barely related one—is often considered evidence of its passivity (Irving, 2016). Yet mind-wandering also seems active in certain respects. It is widely recognized to help advance our goals (Carruthers, 2015; Irving, 2016; Sripada, 2018; Shepherd, 2019; Junker & Grünbaum, 2024; Murray, 2025).[1] Explaining this feature gives us the second desideratum:

The Goal Pursuit Thesis: mind-wandering contributes to the advancement of our goals.

Extant work addressing its role in goal pursuit has focused on the causes of goal-related mind-wandering (Carruthers, 2015, Chap. 6; Irving, 2016), or how cost-benefit computations initiate mind-wandering as a means of exploring new, potentially useful information (Kurzban et al., 2013; Sripada, 2018; Shepherd, 2019; Murray, 2025). However, the generation of goal-relevant information alone does not explain how mind-wandering can advance our goals—the information might never be used in goal-advancing behavior. Hence, we must explain how the information is integrated into processes that advance our goals, and what these processes might be. A natural candidate is inference: processes that draw conclusions about what is the case or what to do, including about how to achieve and adjust our goals. The resulting conclusions can guide subsequent reasoning and action, thereby advancing our goals. Yet little has been said about how this integration is achieved. To address this, I propose the following account:

> *The Learning Account:* mind-wandering informs conclusions about what is the case or what to do. This is enabled by mechanisms that monitor, evaluate, and regulate mental representations and processes, under the active guidance of the executive control system.

Such mechanisms are often assumed to be absent during mind-wandering, leading some to suggest that its role in goal pursuit can be reconciled with passivity. The missing agentive features are typically taken to involve forms of monitoring and regulation (Irving, 2016; Murray, 2025). I argue, however, that these very mechanisms are integral to explaining how mind-wandering advances goal pursuit.

The role of executive control is particularly interesting. If mind-wandering is monitored and regulated by the executive control system, then it is an actively guided mental process.

[1] Some also argue that we can intentionally initiate and sustain episodes of mind-wandering (Irving, 2021). I will largely set this issue aside.

This challenges the standard view that mind-wandering is passive and unguided. The tension arises because the role of mind-wandering in goal pursuit has not been analyzed in sufficient detail. Once we identify the mechanisms that explain this role, it becomes apparent that the mechanisms whose absence was thought to explain its passivity are not absent after all. We should therefore let go of the passivity thesis and embrace the active nature of mind-wandering.

The paper proceeds as follows. Section 2 explains how the mechanisms underlying mind-wandering account for its role in goal pursuit and argues that it involves active guidance by the executive system. Section 3 outlines the advantages of the view that mind-wandering is guided and addresses objections. Section 4 defends and develops the claim that mind-wandering is active and discusses features of agency that mind-wandering lacks. Finally, Section 5 argues that mind-wandering is a learning process that informs conclusions and discusses implications for rational inference and responsibility.

2. Mechanisms of Mind-Wandering

Taken together, the evidence strongly indicates that mind-wandering supports goal pursuit. People frequently report that their mind-wandering episodes involve planning, reappraising situations, making decisions, or attempting to solve problems, and have future-oriented, self-relevant, and goal-related content. Roughly half of all mind-wandering is about the future and around a quarter is about our plans or goals (Baird et al., 2011; Stawarczyk et al., 2011, 2013), suggesting a role in anticipating and planning for the future. Prospective mind-wandering has also been found to increase the likelihood that planned activities are completed (Kvavilashvili & Rummel, 2020). Mind-wandering is also associated with enhanced creative problem-solving (Baird et al., 2012; Gable et al., 2019). It recruits neural and cognitive mechanisms that overlap with episodic simulation and executive control (Fox et al., 2015), both of which are integral to decision-making and planning. As a result, mind-wandering helps us not only plan what to cook for dinner but also invent creative new recipes. In what follows, I identify the mechanisms by which mind-wandering achieves this.

2.1. Episodic simulation

Episodic simulation—the mental construction of past, future, or counterfactual events (Schacter et al., 2012)—plays a central role in mind-wandering, which is characterized by the continuous generation of such simulations. Mind-wandering has been found to support many of the same functions as episodic simulation, including planning (Baird et al., 2011;

Stawarczyk et al., 2011, 2013; Junker & Grünbaum, 2024), creative problem-solving (Baird et al., 2012; Gable et al., 2019), memory consolidation (Sripada, 2016; Mills et al., 2018; Mildner & Tamir, 2019), and a search for new and better action opportunities (Sripada, 2018; Shepherd, 2019). Episodic simulation is therefore key to explaining the functions of mind-wandering.

Much is now known about the functions of episodic simulation. Episodic simulations elicit emotional reactions, which help us evaluate the desirability of the simulated events (Gilbert & Wilson, 2007; Bulley & Schacter, 2020). For example, when imagining the outcome of waiting for a larger, later reward, I can experience some of the anticipated pleasure of this outcome. This emotional response can dampen delay discounting—the tendency to devalue outcomes as their temporal distance increases—and lead to the realization that working toward a larger, later reward is better than pursuing a smaller, sooner one (Peters & Büchel, 2010). Episodic simulation thus promotes more far-sighted decision-making. Mind-wandering has similarly been shown to reduce delay discounting (Smallwood et al., 2013).

When we imagine alternative past or future events, we can reflect on and evaluate them to gain various insights: we might realize that our beliefs, predictions, or reasoning about these events were incorrect; identify that certain future choices lead to mutually exclusive outcomes; or assess the strengths and weaknesses of our cognitive abilities in different contexts—such as our capacity for exert self-control (Bulley & Schacter, 2020; Redshaw & Suddendorf, 2020). Consider the example of dietary choices: eating tasty but unhealthy foods offers smaller, immediate rewards, whereas a healthy diet promises larger, delayed rewards. By simulating each scenario, I can reflect on their outcomes, realizing that these possible futures are mutually exclusive and that a healthy diet better aligns with my long-term goals. Recognizing that my self-control might falter, I can pre-commit to my goals by removing tempting snacks or, anticipating potential setbacks, devise a contingency plan—such as enrolling in a fitness class.

Emotional reactions also influence evaluation of choices by triggering feelings of regret or relief. We tend to avoid choices we expect to regret and prefer those we expect to look back upon with relief. When we imagine alternatives to our past choices, we may experience regret or relief depending on whether the imagined alternatives seem better or worse than the choices we actually made. Likewise, we can anticipate regret or relief by imagining the outcomes of future actions and how we might feel about them in hindsight. Crucially, these evaluations involve monitoring whether the imagined scenarios align or conflict with our goals and values. The fact that mind-wandering often centers on goals, plans, problems, or decisions and elicits emotional reactions congruent with its content (Poerio et al., 2013) suggests that mind-wandering engages in similar evaluative simulations.

Mental simulation plays a crucial role in making otherwise opaque information accessible to higher cognitive processes. This includes helping us identify the source of certain behavioral outputs and understanding the justification behind those outputs (Aronowitz & Lombrozo, 2020; Miyazono & Tooming, 2024; Shea, 2024a). By simulating an action, the agent can infer its likely causes and effects, as well as evaluate competing hypotheses. This process enables the agent to assess reasons for and against drawing specific conclusions, leading to more informed judgments and decisions.

Mind-wandering likely serves a similar function in extracting further information from stored representations. Through its varied content and temporally spaced replay of specific episodes, mind-wandering creates optimal conditions for interleaved learning, which helps optimize memory (Sripada, 2016; Mills et al., 2018; Mildner & Tamir, 2019, 2024). Repeatedly replaying specific episodes gradually extracts from these episodic memories—rich in concrete details, imagery, and particular spatiotemporal contexts—more abstract, gist-like semantic memories. In this process, these semantic memories become integrated with semantically related representations, allowing the agent to learn more generalizable insights such as statistical regularities, categories, and causal relationships (Kumaran et al., 2016; Sripada, 2016). For example, replay of past restaurant visits can reveal patterns such as which places consistently offer the best food, service, or ambiance.

2.2. Experience replay

The relationship between mind-wandering and experience replay merits closer examination. Like mind-wandering, experience replay has been linked to both memory consolidation and planning (Ólafsdóttir et al., 2018). It involves the sequential reactivation of neurons coding for specific locations, enabling the simulation of navigational paths through familiar environments. These simulations help integrate information about potential rewards along different routes, thereby informing future decisions and plans (Momennejad et al., 2018; Liu et al., 2021). Simulations during mind-wandering appear to serve a similar purpose. Both processes frequently occur during rest and recruit overlapping brain regions (Fox et al., 2015; Higgins et al., 2021), suggesting a functional and neural link between the two.[2]

Experience replay updates the expected value of actions and outcomes. If a person receives a large reward in a particular state, replaying paths that lead to that state strengthens the

[2] While replay sequences are highly time-compressed (operating at the scale of milliseconds), they can lead to cascades of activity involving episodic simulations and semantic memories (Kaefer et al., 2022)—the kinds of content typically associated with mind-wandering.

expected value of choosing those paths, thus updating the brain's expectations for future decision-making. The more replaying an event is expected to improve future decisions and increase rewards, the more it will be prioritized for replay (Liu et al., 2021). This dovetails with theories suggesting that mind-wandering is prioritized when its expected value exceeds that of focused, goal-directed thinking (Sripada, 2018; Shepherd, 2019). These insights can be naturally integrated: mind-wandering preferentially accesses and simulates events that are expected to improve future decisions.

It is increasingly recognized that the mind is deeply evaluative in nature (Railton, 2017; Haas, 2023; Carruthers, 2025; Sripada, 2025). Mental processes continuously evaluate states of affairs as better or worse, routinely updating estimated values of actions and outcomes when actual reward outcomes diverge from expectations. Evaluation is closely tied to the affective system, where reward and cost signals are linked to positive and negative emotional reactions (Bechara et al., 1997; Railton, 2017). Value learning also occurs when events are imagined or simulated (Momennejad et al., 2018; Liu et al., 2021). This further supports the idea that simulations during mind-wandering facilitate learning and suggests that emotional reactions during mind-wandering function as signals for value learning.

In summary, the wandering mind helps us evaluate states of affairs and update value estimates of actions and outcomes. Mind-wandering can sample from a vast array of possible actions, narrowing down the options we are likely to consider to those deemed both likely and valuable. These evaluations alter what we take ourselves to have reason to do. When mind-wandering indicates that prior intentions are achievable and valuable, it strengthens our reasons to remain committed to them. Conversely, when mind-wandering indicates that our intentions are harder to achieve or less rewarding than expected, we acquire reasons to reconsider whether they still represent the best course of action (Junker & Grünbaum, 2024). In sum, substantial evidence suggests that evaluations occurring during mind-wandering yield considerations that can fruitfully inform inference.

2.3. Foraging in memory

To understand its functions, we must also consider the meandering, exploratory dynamics of mind-wandering. This can be explained as a foraging process through semantic memory (Mildner & Tamir, 2019). Much like animals foraging in space—gathering resources from one patch until diminishing returns make it advantageous to search elsewhere—the mind shifts

between clusters of related items in semantic memory. Effective foraging requires balancing exploitation of familiar patches with exploration of new, potentially higher-yielding ones.

In verbal fluency tasks—where participants are asked to name as many items as possible from a category (e.g., animals)—retrieval typically clusters around patches of semantically related items (e.g., pets, then farm animals, then birds). People tend to stay within a patch until the rate of retrieval falls below a certain threshold, at which point they switch to a new cluster. Strikingly, this threshold approximates the long-term average retrieval rate across all patches, making the timing of the switch close to optimal: participants move on when yields have dropped below the overall average (Hills et al., 2012).

Evidence suggests that mind-wandering follows a similar structure. Thoughts cluster around a given topic, with frequent shifts to new clusters of semantically related items (Sripada & Taxali, 2020; Mildner & Tamir, 2024). Viewed as a foraging process, mind-wandering can be modeled as movement through the landscape of memory: the wandering mind follows associative pathways through a semantic network, often in a stochastic fashion. Because connections are denser within clusters than between them, this random walk is more likely to yield transitions within a topic (e.g., from one farm animal to another) than across topics (e.g., from farm animals to birds). Shifts to new clusters typically occur when the retrieval rate within the current cluster slows sufficiently.

Together, these processes help explain the dynamics of mind-wandering. Random walks promote exploration within a semantic cluster, while foraging provides a mechanism for shifting to new topics once the yield of novel information drops too low. As a result, mind-wandering unfolds as a continuous stream of thoughts, with most transitions staying within a topic but some spanning considerable semantic distances. These transitions are guided not by pure randomness but by the associative structure of semantic memory.

Semantic memories can also trigger episodic simulations of related content. For example, recalling facts or names of animals may elicit a simulation of those animals. In this way, the content of mental simulations is partly shaped by an underlying semantic search. Episodic and goal-relevant details also seem to modulate transitions: as the rate of new details slows, the likelihood of shifts to new topics increases (Mildner & Tamir, 2024).

Crucially, if a central function of mind-wandering is memory foraging, this implies ongoing monitoring of foraging efficiency— specifically, whether the threshold for switching to a new semantic patch has been reached. Evidence for memory foraging thus supports the view that mind-wandering involves monitoring whether current thought patterns meet standards for optimal cognition.

2.4. Guidance of mental activities

Let us now turn to the mechanisms that regulate mind-wandering. Although its content is varied, mind-wandering often gravitates toward certain topics more than others, reflecting the influence of one's current state and context. This influence can be understood as a form of regulation of mental activities or, in philosophical terms, guidance. Guidance may be *passive*, but importantly, it can also be *actively exercised* by the agent.

Affective states play a central role in guiding mind-wandering. Mind-wandering's affective content often mirrors one's current mood, which in turn shapes its temporal orientation. For instance, negative moods tend to bias mind-wandering toward present concerns (Poerio et al., 2013). If I regret a recent social misstep, my thoughts are likely to drift toward themes resonant with that emotion. Guidance by affective states is, in a certain sense, passive in nature: these states capture one's attention and guide activities independently of one's intentions or direct control over one's activities.

More significantly, mind-wandering can also be actively guided. Goal states shape its trajectory: priming participants to think about a future task or goal biases mind-wandering toward that goal (Stawarczyk et al., 2011). For example, someone primed to think about a biology test will tend to mind-wander about biology-related topics. The capacity to guide (mental) activities toward one's goals depends on the executive control system and constitutes active guidance of one's activities as they unfold (Watzl, 2017, Chap. 7; Buehler, 2022).

Evidence supports that mind-wandering engages the executive system. Functional connectivity is observed between executive control regions and the default mode network during mind-wandering (Christoff et al., 2009; Fox et al., 2015; Turnbull et al., 2019). This coupling is implicated in goal-directed simulations for planning and problem-solving—both of which also occur during mind-wandering (Schacter et al., 2012). Moreover, the executive system appears to prioritize mind-wandering strategically when external task demands are low and to decouple attention from external stimuli to facilitate efficient processing of self-generated content (Turnbull et al., 2019).

More specifically, mind-wandering engages core executive functions that are crucial for understanding its role in goal pursuit. These include activating goal-relevant representations; maintaining and manipulating representations in working memory; enhancing the processing of representations to which executive resources are allocated; and inhibiting distractions and prepotent responses (Diamond, 2013; Buehler, 2022; Badre, 2025). All these functions play a role in mind-wandering, although since mind-wandering is typically not related to current

tasks, they operate somewhat differently than when guiding current tasks. Because the goals entertained rarely demand immediate action, attention is free to shift between topics, generating information relevant to multiple goals. Executive control downregulates processing of external stimuli and current-task representations, redirecting resources toward efficient processing of self-generated content. Inhibiting prepotent responses facilitates exploration of a broad range of possibilities. Activation and manipulation of goal-relevant representations in working memory enables the construction of mental simulations about how to achieve or adjust one's goals. Integrating various sensory, motoric, affective, and evaluative representations in working memory to construct such simulations requires executive functions. This process outputs new representations of actions and outcomes that inform reasoning and action (Carruthers, 2015; Shea, 2024a) and explains how mind-wandering supports goal pursuit.

Recent computational theories offer further insights into the role of executive control in mind-wandering. According to the influential expected value of control theory, the executive system continuously evaluates the costs and benefits of deploying different forms of control in different contexts (Shenhav et al., 2017). On this view, mind-wandering is prioritized when its expected value—the anticipated benefits of success, weighted by the likelihood of success and discounted by expected costs—exceeds that of maintaining focus on the current task (Kurzban et al., 2013; Sripada, 2018; Shepherd, 2019; Murray, 2025).

2.5. Meandering

At first glance, guidance may seem to conflict with memory foraging—the former implying constraints, the latter a lack thereof. In the same vein, the meandering dynamics of mind-wandering is sometimes considered evidence that it is passive and unguided (Irving, 2016). Yet guidance and meandering can be reconciled. Meandering can be understood as a form of memory foraging that is not suppressed by guidance but made possible by it.

Firstly, executive control enables a meandering stream of thought by decoupling attention from external stimuli and inhibiting prepotent responses. But its role in facilitating meandering likely extends beyond this. The executive system continuously monitors control-relevant signals such as conflicts between responses, execution errors, uncertainty, and reward (Shenhav et al., 2016; Badre, 2025). An analogous form of monitoring occurs in memory foraging to determine when it is advantageous to abandon a semantic patch and search elsewhere. Executive control is a strong candidate for the mechanism that detects optimal switch points—moments when the expected value of shifting patch outweighs that of staying—

and initiates the switch. On this view, executive control not only removes barriers to memory foraging but actively regulates its dynamics.

The executive system may also learn to strategically prioritize meandering as it learns the value of the learning opportunities it affords. Meandering allows us to explore and evaluate a broad range of possibilities and extract generalizable lessons from past experiences. Curiosity likely plays a motivational role here: acquiring new information (or learning) is rewarding, and curiosity motivates the pursuit of such rewards (Marvin & Shohamy, 2016; Carruthers, 2024). Since meandering thought is a way of acquiring new information, it can likely be motivated by curiosity. Accordingly, when an ongoing task yields diminishing returns or increasing costs, the executive system may deem the expected value of mind-wandering to be higher, actively downregulating task-related processing and shifting resources toward mind-wandering.

This is consistent with mind-wandering's bias toward certain contents. Exploration biased toward goal-relevant information will often be more valuable than wholly unguided exploration. To sustain long-term plans, it is advantageous to consider reasons and means that support current intentions rather than continuously entertain alternative goals (Junker & Grünbaum, 2024). Guidance and exploration, then, are not opposed but complementary.

3. Guidance Redux

Mind-wandering, then, contributes to goal pursuit through mechanisms that monitor, evaluate, and regulate mental representations and processes. In contrast to prevailing accounts that portray it as passive, this view holds that mind-wandering is actively guided. In this section, I outline the advantages of this account and address potential objections.

3.1. Monitoring, evaluation, and regulation are linked

Zachary Irving (2016, 2021) offers an influential account of mind-wandering as unguided attention: when our minds wander, nothing redirects our attention back to a task or goal once it has drifted. On his view, mind-wandering is 'not monitored or regulated in the right way to count as guided' (Irving, 2016, p. 563). Guidance, for Irving, requires a distinctive phenomenology: we monitor for deviations of attention away from current tasks or goals, experience such deviations as distractions, and regulate felt distractions by reorienting attention back to tasks or goals. Mind-wandering, by contrast, is defined by the absence of this phenomenologically mediated monitoring and regulation.

However, as we have seen, it is precisely the monitoring, evaluation, and regulation of mental representations and processes in goal-conducive ways that explain how mind-wandering supports goal pursuit. Goal-related content that is not recognized as such will usually not inform reasoning and action. Without some appraisal of its potential value or relevance, there is no signal to downstream systems about how the information should be used. As we have seen, mind-wandering involves various kinds of evaluative processing, and such evaluations presuppose monitoring. Any evaluation requires a standard: when we appraise a state of affairs as better or worse, we do so relative to some benchmark. In the case of mind-wandering, the benchmark is the content's relevance to our goals and values. We evaluate whether simulated actions and outcomes are desirable and whether they support—or conflict with—our goals. A positive evaluation indicates alignment with our goals and values; a negative one signals conflict. Crucially, evaluations depend on prior monitoring that registers the applicability of the standard: only once we determine the content is relevant to our goals or values can we proceed to evaluate whether it is favorable or unfavorable in light of them.

One might argue that the evaluation needed to advance our goals occurs only when representations are later accessed during reasoning, rather than at the moment they are first generated. On this view, mind-wandering merely encodes and consolidates potentially useful representations without evaluating them in real time. However, evidence shows that relevant forms of evaluation do occur during mind-wandering. Moreover, for subsequent reasoning to effectively retrieve relevant representations, it helps if their value and task-relevance have already been assessed and encoded. Consolidation without evaluation thus struggles to explain how consolidated representations are prone to be accessed at opportune moments to guide reasoning and action.

Evaluation and relevance-monitoring are thus central to explaining mind-wandering's role in goal pursuit. But why also posit active regulation of our mental activities? One might argue that monitoring and evaluation alone suffice. Susanna Siegel (forthcoming) contends that mind-wandering consists of spontaneous inquiry into questions without guidance from deliberate decisions, concerted efforts, or intentional, self-aware actions. During mind-wandering, we evaluate ongoing inquiries by applying standards to which we hold ourselves. According to Siegel, this evaluative activity does not impose pressure to maintain a single point of focus; attention need not be regulated. Questions and answers arise unbidden, and pieces of information are assessed as possible answers—or as favoring or disfavoring certain answers—without pressure to redirect attention to a previous line of thought.

While the wandering mind does engage in inquiry, evidence shows that this process is also regulated in goal-conducive ways. Mind-wandering is biased toward goal-relevant content, strategically prioritized when external demands are low, decouples attention from perception, and recruits executive control networks. Executive functions facilitate exploration and access and integrate representations that can guide reasoning and action, explaining how mind-wandering can form the kinds of representations that support goal pursuit. Taken together, these considerations show that the functional profile of mind-wandering is best explained by mechanisms that monitor, evaluate, and regulate mental representations and processes.

3.2. Task-unrelated thought

It might be argued that active guidance of mind-wandering conflicts with standard scientific practice. According to Samuel Murray (2025), monitoring and regulation of one's task performance conflicts with the prevailing scientific understanding of mind-wandering as task-unrelated thought. To accommodate this, he defines mind-wandering as thought that lacks vigilance—a capacity to manage multiple goals over time. Vigilance involves '(1) monitoring for circumstantial and task-relevant information that, when perceived (2) triggers implementing task-appropriate representations that are (3) maintained through the completion of the task (or task-segment) or until the agent revises their intention.' (ibid., p. 8).

On this view, mind-wandering lacks the monitoring and regulation of task performance that supports temporally extended agency. An advantage of this view is that it explains why mind-wandering is task-unrelated: it lacks the vigilance that monitors and regulates task performance. It thus vindicates the standard operational definition of mind-wandering as task-unrelated thought. This account thereby aligns with the vast majority of empirical studies; any account violating the task-unrelated definition would be inconsistent with much of the literature.

Like on the present account, Murray links monitoring and regulation of task performance to executive control. One might worry, then, that active guidance by the executive system renders mind-wandering task-related. However, active guidance does not conflict with the standard paradigm. In typical studies, episodes are classified as mind-wandering when they are unrelated to participants' *current* tasks. One role of executive control during mind-wandering is precisely to disengage from the ongoing task, enabling a potentially more valuable mode of thought. While this often involves activating goal representations, these typically concern goals *other than* the current task. Thus, active guidance does not make mind-wandering task-related in the relevant sense. On the contrary, it facilitates decoupling from the current task.

3.3. Rumination and goal-directed thinking

Another worry is that if mind-wandering is guided, it may be difficult to distinguish from other constrained forms of thought, such as rumination or goal-directed thinking. However, the account developed here can still distinguish these phenomena. Different factors influence the trajectory of mind-wandering dependent on the individual's current state and context. Mind-wandering tends to range freely, exploring diverse topics. Guidance mechanisms sometimes constrain the search, biasing it in particular directions. Intense affective states can override broad search and fixate attention on emotionally charged topics. Rumination occurs when strongly negative affective states fixate attention and prevent thoughts from drifting elsewhere.

While mind-wandering is often goal-directed, it differs from the focused, goal-directed thinking that guides current tasks. In the latter, executive control sustains attention on what is most relevant to one's current task. By contrast, during mind-wandering, executive control decouples attention from current tasks and facilitates a broader, more exploratory search.

This is broadly consistent with the dynamic framework of thought (Christoff et al., 2016), in which moderate constraints distinguish mind-wandering from other modes of thought: rumination is more constrained by affective states, and goal-directed thinking is more constrained by executive control. One potential difference concerns the role of executive control. Rather than necessarily being subject to modest executive control, the present account suggests that mind-wandering repurposes executive resources: instead of focusing attention and guiding performance of current tasks, the executive system prevents prolonged focus on any single topic and facilitates a broad search, evaluation, and integration of representations.

4. Passivity Reconsidered

We have seen that there is strong reason to think that mind-wandering is, in some sense, guided by certain mechanisms. But why should we consider the exercise of these capacities something we actively do, as opposed to something that merely happens to us? In this section, I defend and elaborate the view that mind-wandering is active.

4.1. Executive failure

One potential objection is that mind-wandering is often an unhelpful distraction or a failure to exert the executive control necessary to sustain focus and succeed with ongoing tasks (McVay & Kane, 2010). But while mind-wandering can interfere with task performance, it generally

reflects an adaptive use of executive resources. We tend to mind-wander strategically when task demands are low. Moreover, individuals with higher working memory capacity mind-wander less when demands are high but more when they are low, suggesting that their extra executive resources allow them to more flexibly regulate mind-wandering in response to situational demands (Levinson et al., 2012; Rummel & Boywitt, 2014).

The executive system's capacity to learn the expected value of various control strategies helps explain its adaptive regulation of mind-wandering (Lieder et al., 2018). By observing the outcomes of mind-wandering, the executive system may learn that using executive resources to regulate mind-wandering sometimes has high expected value, for example, when external demands and the costs of mind-wandering are low, or when the current task has low expected value. This is why people often mind-wander during routine, undemanding activities such as cleaning or showering, or in unrewarding situations like long, unimportant meetings. In such contexts, prioritizing mind-wandering need not be a failure of executive control. Yet it may still feel like a failure of control, because the computations driving this prioritization are largely unconscious (as discussed below), while the negative consequences of mind-wandering—such as inattentiveness to one's surroundings—are often apparent.

Mind-wandering does carry significant costs, including reduced reading comprehension, increased risk of traffic accidents, and poorer performance on external tasks requiring extensive executive functioning (Smallwood & Schooler, 2015). One explanation for cases where mind-wandering is inappropriately prioritized is that the executive system is miscalibrated, leading to faulty cost-benefit assessments. This may stem from insufficient or unrepresentative sampling of the costs and benefits of different control strategies in the current context, causing the system to underestimate the value of sustained focus or the costs of mind-wandering.

The commandeering effects of affective states also help explain why mind-wandering is sometimes maladaptive. Intense affective states—alluring or distressing—can hijack the stream of thought, diverting it away from more constructive pursuits. This can give rise to negative feedback loops. For example, while mind-wandering about plans with friends, one might be reminded of a recent social misstep. Recalling the awkward moment may prompt simulations of how it was perceived, eliciting emotional responses that deepen the sense of regret and further entrench the negative thought pattern. In this way, intense affective states can turn mind-wandering into unhelpful rumination.

4.2. Unconscious guidance

Another concern is that mind-wandering often appears to us as something that merely happens to us as, rather than as an activity we actively initiate or guide. A plausible response is that we are simply unaware of the mechanisms that guide it. In this vein, Peter Carruthers (2015, Chap. 6) proposes that mind-wandering involves unconscious, subpersonal decisions to direct attention to contents activated by associative processes that are appraised as relevant to one's goals and values. When relevant contents attract attention, they can pass the threshold for consciousness and get broadcast to other systems for further processing. According to Carruthers, these attentional shifts are action-like but not full-fledged actions, since they are not carried out by the whole person but by cost-benefit analyses in the anterior cingulate cortex.

While Carruthers' view aligns with mine in recognizing active guidance during mind-wandering, it goes further in claiming that all attentional shifts are active and subpersonal decisions that act as gatekeepers for consciousness. Not all mind-wandering need be guided by executive processes in this way. Some episodes are passively guided. Moreover, attention is arguably better understood as a personal-level activity—especially when executive control is involved (Wu, 2011; Watzl, 2017; Buehler, 2022). When the executive system guides the individual's stream of thought, it coordinates the activities of their parts, integrates information from various subsystems, and compensates for interference. Guiding one's activities in this way is something done by the whole person, not just a particular neural subsystem.

Finally, we need not resolve debates about whether attention is necessary or sufficient for consciousness. To explain the impression of passivity, it suffices to note we are unaware of the mechanisms guiding mind-wandering. While we are unaware of the mechanisms themselves, their output might enter the conscious stream of thought. We need not be aware that monitoring, evaluation, and regulation were involved in its production. Content may also become conscious for reasons unrelated to goals and values—for example, through automatic responses or perceptual and affective salience. The claim that mind-wandering is actively guided is therefore compatible with a broad range of views on attention and consciousness.

4.3. Beyond a simple view of agency

One might challenge active guidance by trying to offer an alternative explanation for mind-wandering's seemingly agentive features. Irving (2016, 2021) argues that beliefs, desires, and goals cause mind-wandering episodes that relate to those states. As he puts it, 'having a *goal* and *believing* that thinking is a means to achieve the goal *causes* one's mind to wander to that

goal' (Irving, 2016, p. 552, original emphasis). However, while goals may cause goal-related mind-wandering, Irving maintains that this does not result in guidance of the episode as it unfolds. On his view, the lack of guidance explains the passivity of mind-wandering.

Yet Irving offers no explanation of how these mental states exert their causal influence, nor how mind-wandering contributes to goal pursuit. A full explanation needs to explain the effects of such states on subsequent mental processing. The present account addresses this gap. When a goal initiates a behavior, additional representations and processes are needed to guide it toward a successful outcome. The same holds for mind-wandering episodes caused by goals. Beliefs, desires, and goals influence mind-wandering, in part, because executive control actively maintains and manipulates goal-relevant representations in working memory to construct simulations of actions and outcomes that inform our conclusions.

While simple simulations—such as experience replay sequences—may occur with minimal executive involvement, more complex simulations depend substantially on executive functions. The capacity to imagine alternative possibilities and appreciate that they are mutually exclusive develops relatively late, likely dependent on sufficient increases in working memory capacity (Redshaw & Suddendorf, 2020). Mind-wandering often involves moderately complex simulations of alternative possibilities. For example, mind-wandering about dinner may cue the goal of buying groceries. As this goal attracts attention, it cues the memory that I also need to pick up a parcel, which prevents me from passing by my usual supermarket. This prompts a search for other supermarkets along the route to the parcel shop and a simulation of the shortest path that includes both stops. Running such a simulation requires executive functions to access relevant representations from memory and other specialized systems and integrate them in working memory (Shea, 2024a). A complete explanation of the causal interactions between mind-wandering and our goals therefore includes active guidance by the executive system.

That said, mind-wandering often lacks features associated with agency. Unlike intentional action, which is organized around a specific goal, mind-wandering lacks a unified structure: it does not pursue an overarching goal (beyond perhaps memory foraging) but drifts between loosely connected topics. It can also distract us from tasks we are actively trying to complete. For example, my supermarket simulation might arise as I am trying to finish a different task. This contrasts with intentional planning since my current intention is not to plan errands but to do something else. Mind-wandering is thus often unintentional in the sense that it conflicts with what we currently intend to do. I leave open whether mind-wandering can be intentional in other ways, such as by intending not to think of anything in particular (Irving, 2021).

It might be argued that even when mind-wandering is caused by an intention to resolve some issue and leads to an intended outcome—say, arriving at a conclusion—this transition is merely accidental or deviant.[3] The outcome may not be caused in the right way for the event to qualify as an intentional action. We may, for example, lack the requisite kind or degree of control over the process to reliably produce intended outcomes through mind-wandering. Intending to decide and thereby causing a mind-wandering episode about this aim does not guarantee that an attempt at reaching a decision will be made. Additional factors, such as an intention to settle the issue *at this moment* or *in a certain way*, may be required for the causal chain to be non-deviant. It seems doubtful that one could genuinely intend to decide *by means of* mind-wandering and succeed in a non-deviant, while still counting as mind-wandering. Intentionally deciding is better understood as intentional deliberation, which we ought to distinguish from mind-wandering (as discussed below).

For those who hold that action requires knowledge of what one is doing, the fact that we are largely unaware of what happens during mind-wandering might suggest a lack of agency. Nevertheless, our capacity to reflect on and report the content of mind-wandering indicates that some degree of awareness is possible, albeit perhaps only retrospectively.

Depending on one's conception of agency, these considerations may support the view that mind-wandering falls short of certain forms of agency. That the executive system can guide mind-wandering in the absence of other agentive features shows that different dimensions of agency come apart. At the same time, the fact that mind-wandering is active at all shows that mental agency is more pervasive than often recognized.

5. Mind-Wandering as a Learning Process

Mind-wandering is a heterogeneous phenomenon. Its manifestation at any given moment depends on the particular configuration and combination of its underlying mechanisms, and which of these exert the greatest influence over its content and trajectory. Mind-wandering may therefore lack a sufficiently stable cluster of properties to constitute a natural kind (Boyd, 1991). A unifying theme, however, is that mind-wandering supports learning: it helps explore a wide array of possibilities, evaluate actions and outcomes, and extract generalizable information from past experiences. Mind-wandering may therefore be seen as a learning

[3] This differs slightly from the standard issue of intentional mind-wandering, which concerns whether one can intentionally initiate and sustain mind-wandering, rather than how intentions might influence its content and functional role.

process. This has important implications for other mental phenomena with which mind-wandering overlaps.

Like mind-wandering, creative thinking relies on an interplay between sematic memory and episodic simulation, which together generate ideas that are evaluated for their novelty and effectiveness (Benedek et al. 2023). Mind-wandering has been associated with the idea-generation phase of creativity (Baird et al., 2012; Girn et al., 2020). The present account clarifies the mechanisms by which mind-wandering produces novel ideas, while also highlighting the role of affective states and simulations in evaluating them. While this suggests that certain forms of evaluation already occur during mind-wandering, assessing the more complex implications of one's ideas likely requires more deliberate, stepwise reasoning.

Mind-wandering is also a plausible mechanism for non-deliberative attitude change. To avoid regress, it must be possible to form beliefs and intentions without deliberation (Arpaly & Schroeder, 2012; Railton, 2017). If every intention required deliberation, and each act of deliberation itself presupposed an intention to deliberate, an infinite regress would result. Mind-wandering offers a solution. Unlike intentional deliberation, mind-wandering allows us to reach conclusions—about what is the case or what to do—without first intending to deliberate. For example, while mind-wandering I might imagine neglecting to prepare for tomorrow's presentation and the resulting embarrassment. This can lead me to form the belief that *I would regret not having rehearsed tomorrow*, and the intention to *practice later today*—without my ever intentionally raising the question of whether I should rehearse.

There is also reason to think that transitions during mind-wandering can sometimes qualify as rational inference.[4] On one view, inferences are motivated by metacognitive feelings, such as feelings of reliability (Shea, 2024b). For example, past experience might have taught me that the inference from the thought *I have not rehearsed my presentation* to *I would regret not having rehearsed tomorrow* is generally reliable. As a result, whenever I consider the former, I am disposed to conclude the latter without consulting further premises. Mind-wandering engages processes—such as imagining suppositional scenarios and evaluating them against our goals and values—that are often associated with inference (Myers, 2021; Shea, 2024a). If the metacognitive appraisals of transitions are properly calibrated to reflect the actual reliability, quality, and costs of those transitions, and if these appraisals regulate transitions to conclusions

[4] For the purposes of this discussion, I adopt a broad notion of inference as responses to premise-states that yield conclusions about what is the case or what to do in a way that is faithful to the content of the premise-states. I do not assume that such responses must be governed by logical rules, nor that the subject must be aware of the fact that they are drawing an inference.

during mind-wandering, we may regard these transitions as rational inferences. Metacognitive feelings tracking the quality and costs of (mental) actions seem to be produced by the executive system when guiding action (Shepherd, 2025). Since mind-wandering is guided by the executive system, this makes it plausible that such metacognitive feelings can guide inferences during mind-wandering. Moreover, many now recognize that imagination can generate justified belief or knowledge when the representations that constrain it are sensitive to real features of the world (Williams, 2021; Myers, 2021; Miyazono & Tooming, 2024). Since mind-wandering employs many of the mechanisms associated with properly constrained imagining, their operation during mind-wandering may also suffice for rational inference.

Finally, if mind-wandering is actively guided by our goals and values and draws conclusions about what is the case or how to act, then we may also bear responsibility for how we mind-wander. One potential source of responsibility concerns content: when our values, goals, or implicit biases shape the scenarios we entertain, the result may be wishful thinking or other unreasonable patterns of thought. For example, someone harboring implicit bias against a particular group might repeatedly imagine stereotyped interactions with its members, thereby reinforcing the bias. Another concern is the timing of mind-wandering: failing to maintain focus in situations where attention is required—such as drifting off instead of listening to a speaker—may be blameworthy, especially if driven by bias or a refusal to grant the speaker due credibility. In high-stakes contexts, such as surgery, lapses of attention due to mind-wandering may amount to negligence. Various factors—fatigue, stress, emotional distress, or developmental and clinical conditions—may mitigate or excuse episodes of mind-wandering.

If mind-wandering involves inference, this further strengthens the case that it is something we actively do and bear responsibility for. But even if one insists that inference occurs only outside of mind-wandering episodes, it remains clear that inferences often rely on information generated during mind-wandering—and that the learning processes producing this information are themselves actively guided.

6. Conclusion

Understanding mind-wandering requires situating it within the broader mental economy and clarifying its underlying mechanisms and functions. I have sought to do so by highlighting several central features. Mind-wandering is continuously monitored, evaluated, and regulated, enabling it to generate useful information that informs conclusions about what is the case or what to do. Active guidance by the executive control system plays a pivotal role, challenging

the view that mind-wandering is passive and unguided. At the same time, the present account accommodates features often explained in terms of passivity and lack of guidance. What emerges is a picture of mind-wandering as an actively guided learning process that informs reasoning and action, often in rational ways. In turn, this has significant implications for our understanding of mental agency, rational inference, and responsibility.

References

Aronowitz, S. & Lombrozo, T. (2020). Learning Through Simulation. *Philosophers' Imprint,* 20.

Arpaly, N. & Schroeder, T. (2012). Deliberation and Acting for Reasons. *Philosophical Review,*

121:209-239.

Badre D. (2025). Cognitive Control. *Annual Review of Psychology*, *76*(1):167–195.

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, *20*:1604–1611.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012).

Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, *23*:1117–1122.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*:1293-1295

Benedek, M., Beaty, R. E., Schacter, D. L., Kenett, Y. N (2023). The role of memory in creative ideation. *Nature Reviews Psychology 2*:246–257.

Boyd, Richard (1991). Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds. *Philosophical Studies, 61*(1):127-148.

Buehler, D. (2022). Agentive capacities: a capacity to guide. *Philosophical Studies, 179*(1):21-47.

Bulley, A., & Schacter, D. L. (2020). Deliberating trade-offs with the future. *Nature Human Behaviour*, *4*(3):238–247.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford University Press.

Carruthers, P. (2024). The Contents and Causes of Curiosity. *British Journal for the Philosophy*

of Science, 75(4):871-895.

Carruthers, P. (2025). *Explaining our Actions: A Critique of Common-Sense Theorizing*. Cambridge University Press.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive control system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106*:8719-8724.

Christoff, K., Irving, Z., Fox, K., Spreng, N., Andrews-Hanna, J. (2016). Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, *17*:718–731.

Diamond A. (2013). Executive functions. *Annual Review of Psychology*, *64*:135–168.

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage*, *111*:611–621.

Gable, S. L., Hopper, E. A., & Schooler, J. W. (2019). When the muses strike: Creative ideas of physicists and writers routinely occur during mind wandering. *Psychological Science, 30*(3):396–404.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science, 317*(5843):1351–1354.

Girn, M., Mills, C., Roseman, L., Carhart-Harris, R. L., & Christoff, K. (2020). Updating the dynamic framework of thought: Creativity and psychedelics. *NeuroImage, 213*:116726.

Haas, J. (2023). The evaluative mind. In J. Haugeland, C. F. Craver, & C. Klein (eds.), *Mind Design III: Philosophy, Psychology, and Artificial Intelligence* (pp. 295-313). MIT Press

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., & Woolrich, M. (2021). Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron*, *109*(5):882–893.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2):431–440.

Irving, Z. C. (2016). Mind-wandering is unguided attention: accounting for the "purposeful" wanderer. *Philosophical Studies, 173*:547-571.

Irving, Z. C. (2021). Drifting and directed minds: The significance of mind-wandering for mental agency. *The Journal of Philosophy*, *118*:614-644.

Junker, F. T., & Grünbaum, T. (2024). Is the wandering mind a planning mind? *Mind & Language*, *39*:706–725.

Kaefer, K., Stella, F., McNaughton, B. L., & Battaglia, F. P. (2022). Replay, the default mode

network and the cascaded memory systems model. *Nature Reviews Neuroscience*, *23*(10):628–640.

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: an experience-sampling study of working memory and executive control in daily life. *Psychological Science*, *18*(7):614–621.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*:932.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning processs do Intelligent Agents Need? Complementary Learning processs Theory Updated. *Trends in Cognitive Sciences*, *20*(7):512–534.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6):661–679.

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: A review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology, 24*(3):210–237.

Levinson, D. B., Smallwood, J., & Davidson, R. J. (2012). The persistence of thought: Evidence

for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science, 23*(4):375–380.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(4).

Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, 372:(6544).

Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology. General*, *145*(3):266–272.

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin*, *136*(2):188–197.

Mildner, J. N., & Tamir, D. I. (2019). Spontaneous Thought as an Unconstrained Memory Process. *Trends in Neurosciences*, *42*(11):763–777.

Mildner, J. N., & Tamir, D. I. (2024). Why do we think? The dynamics of spontaneous thought reveal its functions. *PNAS Nexus*, *3*(6).

Mills, C., Herrera-Bennett, A., Faber, M., Christoff, K. (2018) Why the Mind Wanders: How

Spontaneous Thought's Default Variability May Support Episodic Efficiency and Semantic Optimization. In K. C. Fox, & K. Christoff (eds.), *Oxford Handbook of Spontaneous Thought and Creativity*. Oxford University Press.

Miyazono, K. & Tooming, U. (2024). Imagination as a generative source of justification. *Noûs*, *58*(2):386-408.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*:e32548.

Murray, S. (2025). Vigilance and mind wandering. *Mind & Language*, *40*(2):174-194.

Myers, J. (2021). Reasoning with Imagination. In Amy Kind & Christopher Badura (eds.), *Epistemic Uses of Imagination* (pp. 103-121). Routledge.

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology, 28*(1):R37–R50.

Peters, J., & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*(1):138–148.

Poerio, G. L., Totterdell, P., & Miles, E. (2013). Mind-wandering and negative mood: does one thing really lead to another?. *Consciousness and Cognition*, *22*(4):1412–1421.

Railton, P. (2017). At the Core of Our Capacity to Act for a Reason: The Affective System and Evaluative Model-Based Learning and Control. *Emotion Review*, *9*(4):335-342.

Redshaw, J., & Suddendorf, T. (2020). Temporal Junctures in the Mind. *Trends in Cognitive Sciences*, *24*(1):52–64.

Rummel, J., & Boywitt, C. D. (2014). Controlling the stream of thought: working memory capacity predicts adjustment of mind-wandering to situational demands. *Psychonomic Bulletin & Review*, *21*(5):1309–1315.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron*, *76*(4):677–694.

Shea, N. (2024a). *Concepts at the Interface*. Oxford University Press.

Shea, N. (2024b). Metacognition of Inferential Transitions. *The Journal of Philosophy*, *121*:(11):597-627.

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the

value of control. *Nature Neuroscience*, *19*(10):1286–1291.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M.

M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*:99–124.

Shepherd, J. (2019). Why does the mind wander? *Neuroscience of Consciousness, 2019*(1).

Shepherd, J. (2025). Salvaging the "sense of agency": Metacognitive feelings for flexible behavioral control. *The Journal of Philosophy*.

Siegel, S. (forthcoming). Wandering Inquiry. *The Journal of Philosophy.*

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*:487–518.

Smallwood, J., Ruby, F. J., & Singer, T. (2013). Letting go of the present: Mind-wandering is associated with reduced delay discounting. *Consciousness and Cognition, 22*(1):1–7.

Sripada, C. (2016). Imaginative Guidance: A Mind Forever Wandering. In M. E. P. Seligman, P. Railton, R. Baumeister, & C. Sripada (eds.), *Homo Prospectus* (pp. 103-131). Oxford University Press.

Sripada, C. (2018). An exploration/exploitation tradeoff between mind wandering and goal-directed thinking. In K. C. Fox, & K. Christoff (eds.), *Oxford Handbook of Spontaneous Thought and Creativity*. Oxford University Press.

Sripada, C. (2025). The valuationist model of human agent architecture. *Philosophical Psychology*, 1–30.

Sripada, C., & Taxali, A. (2020). Structure in the stream of consciousness: Evidence from a verbalized thought protocol and automated text analytic methods. *Consciousness and Cognition*, *85*:103007.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica*, 136(3):370–381.

Stawarczyk, D., Cassol, H., & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology*, *4*:425.

Turnbull, A., Wang, H. T., Murphy, C., Ho, N. S. P., Wang, X., Sormaz, M., Karapanagiotidis, T., Leech, R. M., Bernhardt, B., Margulies, D. S., Vatansever, D., Jefferies, E., & Smallwood, J. (2019). Left dorsolateral prefrontal cortex supports context-dependent prioritisation of off-task thought. *Nature Communications*, *10*(1).

Watzl, S. (2017). *Structuring Mind. The Nature of Attention and How it Shapes Consciousness*. Oxford University Press.

Williams, D. (2021). Imaginative Constraints and Generative Models. *Australasian Journal of Philosophy, 99*(1):68-82.

Wu, W. (2011). Attention as Selection for Action. In C. Mole, D. Smithies & W. Wu (eds.) *Attention: Philosophical and Psychological Essays* (pp. 97-116). Oxford University Press.

ARTICLE 3

Reasoning With Cognitive Maps

Abstract: Cognitive maps are mental representations of geometric structures that (often) lack logical and propositional structure. In this paper, I demonstrate that content-specific mental transitions mediated by cognitive maps can satisfy common conditions for reasoning: conclusions arise as responses to premise-states; transitions are responsive to rational norms; and the reasoner takes their conclusion to be supported by preceding states and operations on which they base their conclusion. This challenges the view that reasoning exclusively takes the form of rule-governed operations over propositional attitudes. Instead, we should acknowledge that a wider range of representational structures and operations can support reasoning.

1. Introduction

Different representational structures have different advantages and limitations. Language-like representations are great for representing logical and propositional structure, whereas maps are great for representing geometric structures. However, debate continues over how to characterize distinct representational structures and the specific operations supported by each.

This paper sets out to examine the nature of cognitive maps—mental representations of geometric structures—and their role in reasoning. While mostly studied in the context of spatial navigation, cognitive maps are increasingly found to be involved in processes commonly associated with reasoning. In this paper, I argue that a clearer understanding of the functional profile of cognitive maps challenges traditional rule-following accounts of reasoning. Conditions for reasoning are met in processes mediated by cognitive maps: conclusions arise as responses to premise-states; the transitions involved are responsive to rational norms; and the reasoner takes their conclusion to be supported by the states and operations on which they base their conclusion. These transitions, however, do not perform the rule-governed operation over propositional attitudes often thought to constitute reasoning.

The paper proceeds as follows. In Section 2, I introduce cognitive maps and their representational mechanisms and properties. In Section 3, I discuss influential accounts of reasoning and present some general, commonly endorsed conditions for reasoning. In Section 4, I show how content-specific transitions mediated by cognitive maps meet these conditions.

2. Cognitive Maps

Cognitive maps have been studied most extensively in the context of spatial navigation. To solve certain navigational tasks, the agent constructs a map of the geometric features of physical space, such as metric (e.g., distances and angles) or topological (e.g., connectedness and adjacency) relations. This is done by combining information about one's own position, based on a running record of self-motion and perception-based estimates of one's spatial relations to local landmarks, with position estimates of previously encountered landmarks. Once equipped with such a map, the agent can compute the path to a goal even in the absence of sensory cues about the target's location or the agent's progress toward it. This capacity extends to situations where the starting location, destination, or path is novel to the agent. (Rescorla, 2017; Whittington et al., 2020, 2022).

A mental representation can be understood as a cognitive map in both a loose and a strict sense. In the *loose sense,* it represents geometric aspects of the environment without itself exhibiting geometric structure. In the *strict sense*, a cognitive map has the same representational properties and mechanisms as concrete physical maps: its constituents and their relations are themselves geometrically structured in a way that corresponds to the structure of the entities and relations the map represents. Both the representational content and format are geometrically structured in a structure-preserving manner (Rescorla, 2009). I will focus on cognitive maps in the structure-preserving sense, while remaining open to differences in the representational mechanisms and properties of concrete physical maps and cognitive maps.

Much is now known about how the brain constructs cognitive maps. Several distinct neuronal cell types have been discovered that encode geometric relations. These include hippocampal place cells, which respond selectively to particular spatial locations, and entorhinal grid cells, which respond to multiple locations arranged in a grid-like activation pattern. Other cell types forming the neural substrate of cognitive maps respond selectively to head direction or the locations of objects, borders, landmarks, or goals. Patterns of activity across these cells carry geometric information. For example, when two place cells consistently fire in rapid succession, their joint activation represents that two locations are spatially close (Whittington et al., 2020, 2022).

In philosophical work on map-like representations, a particular concern has been the extent to which map-like representations exhibit compositionality. It has been argued that systematic relations in the content of maps are reflected in the structure of the thinker's map-like representations and reasoning abilities. Maps are composed of recurring components which

can be flexibly recombined to represent different states of affairs. For example, a blue blob placed in one quadrant of two intersecting lines can be used to represent a lake at a road intersection. The same blob placed between two parallel lines can be used to represent a lake between two parallel roads (Camp, 2007).

Crucially, cognitive maps possess these representational abilities without having to employ the compositional mechanisms of formal logic: predication, logical operators, and quantifiers (Camp, 2007; Rescorla, 2009). Computational models of the use of map-like representations in guiding spatial navigation and visual attention support this view, showing that computations over these representations make use of geometric rather than logical structure (Rescorla, 2009; Buehler, 2025). In this respect, they differ from classical accounts of the compositionality of thought, emphasizing the need for language-like representations and logical compositional mechanisms (Fodor, 1987; Fodor & Pylyshyn, 1988).

Maps also (typically) lack propositional structure. Propositional structures are often characterized by discrete constituents, the ability to combine a wide range of elements, asymmetry (with some elements serving as inputs to others), and recursion—allowing outputs to function as inputs to the same operation. Recursion enables hierarchies of nested iterations of the same representation, as in the sentence 'she is my mother's mother's mother'. Maps, by contrast, are holistic—any marker on a map is automatically related to all other markers—restricted to geometric relations, and non-hierarchical, since the same operation cannot be iterated to form hierarchies. Holism marks a crucial contrast with propositional structures. In propositional systems, updates can be made piecemeal, allowing inconsistencies to persist when new information is not integrated with related propositions. For example, one might believe that penguins are birds and that they cannot fly, yet also believe that all birds can fly—simply by failing to recognize that penguins' inability to fly serves as a counterexample to the belief that all birds can fly. By contrast, updating the location of a marker on a map automatically updates its geometric relations to all other markers (Camp, 2018).

Despite these differences, maps do have veridicality conditions: they are veridical when they correctly represent geometric relations between entities, and non-veridical when not. This need not make them propositional, however. Propositional structures have discrete truth conditions: they are evaluable as true or false. Maps, by contrast, (often) have graded accuracy conditions: they more or less accurately represent geometric relations between entities. For

example, the closer the map's representation of some distance is to the actual distance between entities, the more accurate the representation (Burge, 2018).[1]

It is debated whether maps can be extended with mechanisms like symbolic markers or color coding that function as predicates, logical operators, or quantifiers. But even if we grant such extensions, maps are often not the best way to encode logical information. Language-like systems can readily represent an abstract state of affairs, without committing to the specific concrete facts that instantiate it. By contrast, a map can represent an abstract state of affairs only by representing the locations of the underlying objects and the properties that instantiate it. Since maps represent determinate entities at specific locations, quantificational information is a particularly poor fit for maps. It is hard to represent that *something, somewhere, is F* (e.g., *someone, somewhere, is drinking coffee*). Similarly, since maps only represent a circumscribed area, it is hard to 'fit' universal information such as *all Fs are G* onto a map (Camp, 2007). These types of information are more naturally represented by representational structures with inherent logical and propositional structure—paradigmatically, language-like representations.

Other logical constructs—predicates, conditionals, disjuncts, and negations—are only marginally less awkward to represent on maps. For example, representing the conditional proposition that *if Asha wants a cup of coffee, then she will go to the café* on a map would require introducing mechanisms to represent Asha's desire for coffee, Asha's trip to the café, and the fact that the latter is conditional on the former. We might introduce a coffee symbol on top of Asha to predicate a desire for coffee, color Asha and the coffee symbol red to indicate that this predicate-argument compound is the antecedent of a material implication, draw a line to the café to represent Asha's trip to the café, and color the line blue to indicate that it is the consequent of a material implication.

However, this is an inefficient way of representing such information. It involves first constructing a map on which the relevant entities are located, and then introducing further mechanisms to represent more abstract properties. So as not to confuse antecedent-consequent pairs, new color pairs would have to be introduced for each conditional, and a list would have to be kept over which colors are paired together. The same information could instead be represented in a (language-like) representational structure more naturally suited for logical and

[1] While maps and propositional structures often correlate with distinct veridicality conditions, there may be no sharp distinction. Topological maps representing certain metro stations as connected might be considered true or false depending on whether the connections obtain or not. Likewise, complex collections of sentences, such as books, might be assessed for their overall accuracy. Still, given the discrete, digital nature of propositional structures, it is generally natural to evaluate them in terms of truth rather than accuracy. By contrast, the continuous metric relations on maps, such as distances and angles, are naturally evaluated in terms of accuracy.

propositional content, without the need to locate the constituents on a map. For example, unlike color pairs on a map, words like 'if' and 'then' or the symbol '→' function as universal compositional mechanisms that can combine a wide range of elements, have a fixed meaning, and can be reused indefinitely. This way, the abstract state of affairs (the conditional proposition) could be represented without the extra steps of locating entities and properties on a map. Moreover, while we may conceive of mechanisms like color coding or symbolic markers to extend concrete physical maps, it is unclear that corresponding mechanisms exist for extending cognitive maps in the mind/brain. For these and other reasons, I will focus on cognitive maps without logical or propositional structure.

That said, language-like, propositional thoughts often activate a rich array of other representational structures—including cognitive maps—whose contents relate to the initial thought but also provide additional information (Shea, 2024a). For example, the thought that *if Asha wants a cup of coffee, then she will go to the café* could activate sensory, evaluative or other representations of Asha, the coffee, and the café, as well as map-like representations of their locations and the geometric relations between them. Since different representational structures have distinct advantages and limitations, achieving the best results often depends on deploying them in the right combination—a theme that will become important later.

A strength of cognitive maps is their capacity to flexibly represent and automatically update geometric relations between entities, which supports efficient computation of paths connecting these entities. Representing spatial information in sentences is comparatively inefficient: as the number of entities grows, providing exhaustive representations of geometric relations becomes costly. Sentences often leave many geometric relations implicit, making it effortful to infer additional relations from what is explicitly represented. Consider the following sentences:

Asha is at the café
Bengi is at the library
The park is west of the café
The library is west of the park

From these sentences, we can infer new geometric relations beyond those listed but only with much effort. For example, using the park as a common reference point, we can derive the geometric relations between Asha and Bengi through something like the following steps:

Asha is at the café

The park is west of the café

This means the café is east of the park

So, Asha is east of the park

Bengi is at the library

The library is west of the park

So, Bengi is west of the park

If Asha is east of the park, and Bengi is west of the park, then Bengi is west of Asha

So, Bengi is west of Asha

By contrast, all geometric relations between entities can be read off directly when represented on a map. By locating Asha, Bengi, the café, the library, and the park on a map, it becomes immediately apparent that Bengi is west of Asha. This is a more efficient way of reaching the same conclusion. Reasoning over language-like representations imposes a high cognitive load because each inferential step requires maintaining multiple propositions in working memory, keeping track of intermediate conclusions, and ensuring logical consistency, with errors potentially propagating if any step is fallacious or misremembered. By contrast, representing entities on a cognitive map allows parallel processing: all locations and relations can be represented simultaneously, and the conclusion that Bengi is west of Asha can be read off directly from the map, without the need for step-by-step reasoning or maintaining multiple propositions in working memory. This substantially reduces cognitive load and lowers the likelihood of errors compared to stepwise symbolic reasoning. Cognitive maps scale efficiently: representing further entities and relations simply involves placing new entities on the map, whereas symbolic reasoning might require several inferential steps.

2.1. Empirical findings

Evidence suggests that cognitive maps contribute to a variety of tasks, including some that resemble reasoning. One line of evidence comes from experience replay, in which neural patterns are sequentially reactivated to reinstate past experiences, particularly spatial trajectories. This process engages the neural substrates of cognitive maps, including the hippocampal–entorhinal system, place cells, and grid cells, and supports planning, learning, and memory consolidation (Momennejad et al., 2018; Ólafsdóttir et al., 2018; Wimmer et al., 2023). Importantly, replay facilitates generalization by reorganizing new experiences in sequences consistent with prior knowledge, allowing the integration of new experiences with

current knowledge (Liu et al., 2019; Kurth-Nelson et al., 2023). Replay also aids in learning the reward structure of an environment by rehearsing the trajectories that have tended to lead to rewards (Liu et al., 2021; Wimmer et al., 2023). For example, replay of past restaurant visits can extract patterns—which places consistently serve great food, provide good service, or stay open late—guiding where to go for dinner next.

Cognitive maps can also represent relations beyond spatial ones. One study found that participants constructed a two-dimensional map of a social hierarchy to represent the relative competence and popularity of individuals and identify pairs of individuals that achieved the best combination of the two traits (Park et al., 2021). A key neural substrate of cognitive maps is grid cell activation, which forms a grid over the current environment. The study found an activation pattern in brain areas containing grid cells (entorhinal cortex and medial prefrontal cortex), suggesting that relations between entities were encoded using a grid-like code. A related study found that Euclidean distances between individuals on the social map correlated with reaction times and activation levels in the same brain areas, providing further evidence that relations between individuals were represented using a cognitive map (Park et al., 2020).

Recent theoretical work has also highlighted the role of cognitive maps in conceptual and compositional thought (Bellmund et al., 2018; Frankland & Greene, 2020; Whittington et al., 2020, 2022). An influential proposal by Whittington et al. (2020) suggests that the brain organizes much information, not as isolated facts, but as structured relationships between entities, applicable to both spatial and abstract tasks. Representations of entities and representations of structures in which the entities can be embedded are separated and can be flexibly combined to represent novel states of affairs. The capacity to flexibly combine representations into novel structures supports generalization: structural representations learned in one task can be reused in a different but structurally similar one. For example, the map of a social hierarchy can be repurposed by binding different individuals to locations on the map.

3. What is Reasoning?

What the discussion above suggests is that cognitive maps help us reach conclusions about what is the case (e.g., how entities relate to each other) or what to do (e.g., how to get from one location to another). Since such conclusions are the standard outputs of reasoning, this is at least prima facie evidence that cognitive maps can support reasoning. Before addressing specific accounts of reasoning, let me clarify some terminology. Theorists often use the terms *inference* and *reasoning* interchangeably. I will follow this convention.

One prominent account holds that reasoning a rule-governed operation over propositional attitudes (or their contents) that results in attitude change such as forming a new belief or intention (Broome, 2013; Boghossian, 2014). The rules ensure consistency among propositional attitudes and are broadly logical, employing operators such as NOT, AND, and IF-THEN. This applies to the rules attributed to both theoretical rationality (what it is rational to believe**)** and practical rationality (what it is rational to do), though each domain has its own set of rules. Paradigmatic rules include:

Modus Ponens: If you believe that *if P, then Q*, and believe *P*, then you ought to believe *Q*
Non-Contradiction: You ought not to believe both *P* and *not-P* at the same time
Means-End Coherence: If you intend an end, and you believe that certain means are necessary to achieve it, you ought also to intend those means
Enkratic Rule: If you believe you ought to do something, then you ought to intend to do it

This view restricts reasoning to representations with logical and propositional structure over which the relevant rules can operate. I will refer to this as the *rule-following account*.[2] Other formal rules—such as Bayes' rule or decision-theoretic rules prescribing that the agent selects the option with the highest expected value—do not presuppose representations with logical and propositional structure. As such, they go beyond the rule-following account as characterized here. We will later see that map-mediated reasoning is compatible with rules of this sort.

Language-like systems have the requisite compositional semantics to implement broadly-logical rule-following (Fodor, 1987; Fodor & Pylyshyn, 1988; see also Broome, 2013, p. 267). Cognitive maps, by contrast, (often) lack logical and propositional structure. Yet there is prima facie evidence that they nonetheless support reasoning: they help us draw conclusions about what is the case and about what to do. This gives rise to an inconsistent triad:

Reasoning is a rule-governed operation over propositional attitudes (or their contents)
Reasoning can take place over cognitive maps
Cognitive maps lack logical and propositional structure

[2] Some further details separate proponents of the rule-following account. Both Boghossian and Broome describe reasoning as a personal-level, intentional operation governed by rules. Boghossian (2014) further holds that the reasoner must take the premises to support the conclusion, suggesting that she must be able to become aware of this. By contrast, Broome (2013) does not require such awareness. Moreover, while Boghossian focuses on theoretical reasoning, Broome's account encompasses both theoretical and practical reasoning. For my purposes these distinctions are not central. I will use the term *rule-following* in a broad sense, encompassing both conscious and explicit application of rules as well as cases where rules are unconsciously and implicitly instantiated in mental transitions between attitudes (cf. Quilty-Dunn & Mandelbaum, 2018).

Which claim one rejects determines one's stance on reasoning. Denying the second claim may be called the *exclusionary strategy*, as it rules out cognitive maps as vehicles for reasoning. Denying the third claim might be termed the *subsumptive strategy*: by allowing that cognitive maps can possess logical and propositional structure, it makes room for subsuming operations over cognitive maps under rule-following. Finally, denying the first claim might be called the *liberal strategy*: it adopts a less restrictive view of reasoning—one that may allow reasoning over non-logical, non-propositional representational structures such as cognitive maps.

The key difference between the first two strategies and the third lies in whether one accepts the rule-following account or adopts an alternative conception of reasoning. One might also reject both the first and third claims, combining the subsumptive and liberal strategies. On this hybrid view, reasoning with cognitive maps could be partly subsumed under the rule-following account when they possess the requisite logical and propositional structure, while also allowing that cognitive maps may support reasoning independently of rule-following.

In what follows, I argue against the exclusionary strategy and in favor of the liberal strategy, while leaving open the prospects of the subsumptive strategy. I do not deny that reasoning can occur as described by the rule-following account. Rather, my claim is that this is not the only form of reasoning. Cognitive maps can support reasoning even when they lack logical and propositional structure. So, if not rule-following, what does reasoning amount to? To a first approximation, any reasoning process should satisfy the following conditions:

> *Response Condition*: Conclusions are formed in response to premise-states
> *Rationality Condition*: Transitions are responsive to rational norms

Some version of these conditions appears in most accounts of reasoning, though interpretations vary. On the rule-following account, premise- and conclusion-states are propositional attitudes, and reasoning consists in inferring the conclusion from the premises by following a rule. Responsiveness to rational norms likewise stems from following appropriate rational rules. Some authors also endorse a further condition, influentially formulated by Boghossian (2014):

> *Taking Condition*: Inferring necessarily involves the thinker *taking* their premises to support their conclusion and drawing their conclusion *because* of that fact (p. 5; original emphasis)

For Boghossian, the act of *taking* one's premises to support a conclusion is not a separate mental step but is inherent in the very act of applying a rule. When you use the modus ponens rule to deduce *Q* from *P* and *if P, then Q*, your following the rule simply *is* what it means for

you take the premises to support the conclusion. The taking condition is often interpreted to mean that the reasoner is aware, or can become aware, of what she takes to support the conclusion. Many have pointed out that restricting reasoning to cases where agents consciously and explicitly follow rules overintellectualizes the phenomenon. In response, some deny that taking is a necessary condition for reasoning (Siegel, 2019; Levy, 2024). Others interpret taking in ways that do not presuppose rule-following (Buckner, 2019; Munroe, 2021; Shea, 2024b). Why accept these conditions? They distinguish inferential transitions from merely causal ones where mental states are produced by causal mechanisms that may mimic reasoning, but where no actual reasoning takes place. To borrow an example from Broome (2013):

It is raining
If it is raining, then the snow will melt
The snow will melt

This sequence of thoughts could unfold in a merely causal way, where the final thought is not the right kind of response to qualify as an inference. Thoughts of salt might bring to mind pepper simply because the concepts SALT and PEPPER are associated, but this transition is not an inference. By contrast, inferential transitions engage with the *content* of preceding states and are faithful to that content. If we treat the present case as an inference, the premises refer to entities (rain and snow) and attribute properties to them (occurring, melting). The second premise also introduces logical content in the form of a conditional relation: it specifies how the content of the antecedent (rain occurring) is connected to that of the consequent (snow melting). This transition is faithful to content, and in this deductive case, that faithfulness depends on logical content and takes the form of truth-preservation: if the premises are true, the conclusion must be true. Concluding that the snow will melt is thus an appropriate response to the content of the premises, particularly to the conditional relation.

There is ongoing debate over whether certain formulations of the response and rationality conditions alone suffice for genuine reasoning, or whether some version of the taking condition is also required. While I remain neutral on the necessity of taking, I will argue that reasoning with cognitive maps can involve a form of taking.

3.1. Heterogeneous inference

Why reject the exclusionary strategy? The view that reasoning operates exclusively over logical and propositional structures has come under increasing strain in recent years, as critics have shown that other representational structures can also support reasoning. Sophisticated

forms of associative learning (Buckner, 2019) and operations over iconically formatted mental models and visuospatial imagery (Munroe, 2021) have been argued to qualify as reasoning insofar as they satisfy conditions analogous to the response, rationality, and taking conditions.

A similar case has been made regarding operations over maps. Map-like representations frequently feature in so-called *heterogeneous inferences* that mix representational formats (Aguilera, 2021; Williams, 2024). Such inferences require that representational contents must be sufficiently similar across representations over the steps of the inference. If the entities and attributed properties differ significantly across representations, then mental processes cannot make content-faithful transitions between them. When the premises and conclusion of a reasoning process fail to refer to the same entities and properties—that is, when transitions are not faithful to content—the premises cannot support the conclusion.

Fortunately, content is often preserved across representational formats: they refer to many of the same entities and attribute many of the same properties and relations. For example, an agent might combine a language-like representation of how metro fees increase when crossing zones with a map-like representation of metro lines, stations, and zones to reach the conclusion that the journey from her current location to her destination crosses from one zone to another, and is therefore liable to a higher fee (Aguilera, 2021). Similarly, one can combine a causal graph representing the hypothesis that the noise from inside a cave was caused by a certain individual with a map representing the same individual as standing outside the cave to reach the conclusion that the hypothesis is false (Williams, 2024). In each case, maps are integrated with other types of representation to arrive at a conclusion. This is possible because, despite differences in format, the representations refer to the same entities and attribute the same properties, thus preserving content across transitions.

These proposals rightly recognize that the mind employs a variety of different representational structures—often in combination and suitable for different purposes—and that this diversity also applies to reasoning. My aim is to expand on this insight and show in virtue of what such processes count as reasoning.

3.2. Content-general and content-specific transitions

To bring reasoning across representational formats into clearer view, it is useful to consider Nicholas Shea's (2024a) distinction between content-general and content-specific transitions:

Content-Specific Transition: a transition between representations such that whether or not the transition is faithful to content depends on the content of representations other than broadly-logical terms.

Content-General Transition: a transition between representations such that whether or not the transition is faithful to content depends at most on the content of broadly-logical terms (p. 66).

Deductive inferences, such as modus ponens, are paradigms of content-general transitions:

If Socrates is human, then Socrates is mortal
Socrates is human
Therefore, Socrates is mortal

In such transitions, non-logical terms can be freely substituted (e.g., replacing *Socrates* with *Hypatia*), and the transitions will remain truth-preserving. Faithfulness to content turns solely on the content of logical terms, not on the content of non-logical terms. While practical reasoning is not deductively valid—it is non-truth-preserving, since it operates over non-truth-apt desires and intentions in addition to truth-apt beliefs—standard practical inferences also occupy the content-general end of the spectrum, given their use of broadly-logical terms (e.g., IF–THEN) and rule-governed structure. Consider this example of instrumental reasoning:

I intend to catch the train
If I intend to catch the train, then I must leave now
Therefore, I shall leave now

This reasoning combines an intended end with a belief about the means needed to achieve it, producing an intention to pursue those means. Transitions occur over propositional attitudes, one of which has a conditional proposition as its content, and is guided by the means–end coherence rule. It follows the schema:

I intend to *E*
If I intend to *E*, then I must do *M*
Therefore, I shall do *M*

In this case, faithfulness to content does not depend on the non-logical terms, *E* and *M* (which can be freely substituted), but on the logical structure of broadly logical terms. Such transitions therefore fall toward the content-general end of the spectrum. Now consider a different type of transition. While trying to determine facts about Cyrus, you undergo the following transition:

Cyrus is a dog

Cyrus barks

The latter representation is likely to be true if the former is. In this case, however, faithfulness to content depends on the content of non-logical terms: *Cyrus, dog, barks*. The transition's reliability stems from its sensitivity to the observed regularity that *dogs bark*, even though this regularity is not explicitly represented. Over time, as such *direct* content-specific transitions demonstrate their usefulness (e.g. reliably leading to true conclusions), we are likely to develop dispositions to make them.

Other content-specific transitions between thoughts, Shea argues, are not direct but *mediated* by special-purpose representations (e.g., sensory, motoric, affective, evaluative, or map-like). For example, when trying to determine whether a chair will fit in your car, posing the question may prompt you to imagine rotating the chair to see if some angle allows it to fit. This mental simulation might lead you to conclude that it will probably fit with the back seat down, and, in turn, to form the intention to buy it. Such content-specific transitions lack the logical, rule-governed structure that the rule-following account requires for a process to count as reasoning. I will argue, however, that content-specific transitions mediated by cognitive maps can nonetheless satisfy conditions for reasoning.

4. Map-Mediated Reasoning

Before turning to content-specific transitions, let us first examine the prospects of content-general reasoning with maps. One proposal models route planning with cognitive maps as a form of instrumental reasoning (Aguilera, 2025):

I intend to go from A to B

If I intend to go from A to B, then I must follow the map's indications

Therefore, I shall follow the map's indications

These transitions seem to occur over logically and propositionally structured attitudes, guided by a means–end coherence rule. Map indications are identified and adopted as means to an end. Yet it is not clear that cognitive maps are doing any inferential work in the sense of ensuring faithfulness to content. The reasoning at issue is content-general: it is the content of broadly-logical terms that make transitions faithful to content. The map indications could be replaced by other means, and transitions would remain faithful to content. If cognitive maps are non-logical and non-propositional, they cannot function as broadly-logical terms, nor as logically

and propositionally structured attitudes. For this process to qualify as reasoning *with* rather than merely *about* cognitive maps requires extending cognitive maps with logical and propositional mechanisms. Doing so opens the door for subsuming operations over cognitive maps under content-general reasoning. While I do not rule out this possibility, I aim to show that cognitive maps can facilitate reasoning without such extensions.

4.1. Questioning attitudes

There are other types of processes where cognitive maps can play a substantial inferential role—those where cognitive maps mediate content-specific transitions. These transitions need not solely be between propositional attitudes. As in the example of determining whether a chair will fit in one's car, mediated content-specific transitions can be initiated and guided by *questioning attitudes* (Friedman, 2013; Carruthers, 2018). Examples of such attitudes include curiosity, wondering, and inquiry. These attitudes have not propositions, but questions as their content: *What is that thing? What is over there? Where is home from here?* They presuppose ignorance about a subject matter and are satisfied when relevant information is acquired. Rather than specifying truth-conditions, questions specify the conditions needed to answer them. The question *Where is home?* specifies that a state with the content *home is at location P* is needed to answer the question. What mediates the relationship between the questioning attitude and the answers that satisfy them is learning. Questioning attitudes motivate agents to seek answers to the questions that are their contents. In other words, they motivate learning behavior. This includes actions like approaching unfamiliar objects to investigate them, but also mental actions such as guiding attention and searching memory (Carruthers, 2018, 2025).

This provides a framework for how cognitive maps can mediate content-specific transitions. For example, a questioning attitude with the content *How do I get from A to B?* can initiate and guide a search of one's memory. If I am already highly familiar with the routes connecting these locations, I might already have a map of my preferred route stored and can transition directly to a representation of a route that gets me from A to B. Often, however, we need to plan novel routes in less familiar environments by constructing and manipulating cognitive maps: a cognitive map of the area is retrieved from memory, potential routes are constructed over the map, and the best one is selected, taking into account constraints such as minimizing travel distance. Thus, through manipulations of the map as a mediator, I transition from a questioning attitude with the content *How do I get from A to B?* to a state with the content *this route gets me from A to B*, thereby satisfying my questioning attitude.

This framework can be extended to practical reasoning. When the state representing the route from A to B is paired with an affective state (desire) that (non-propositionally) represents the action or its outcome as valuable, and it is appraised as having higher expected value than relevant alternatives, a further inferential step forms an intention to follow the route (Carruthers, 2025). We will return to this point below.

Transitions from questioning attitudes to answers are content-specific because faithfulness to content depends not on broadly-logical terms, but on responding faithfully to content of non-logical terms (locations A and B and the relations between them) across transitions. Crucially, this satisfies the response condition. The questioning attitude and the sequence of map-like representations elicited by the ensuing search as premise-states can be treated as premise-states, to which we respond with a conclusion that resolves the initial question. While the terms *premise* and *conclusion* are not used in a strict logical sense, the process captures the spirit, if not the letter, of the response condition. To distinguish it from content-general reasoning, we might instead describe the states being responded to as *states of inquiry* or *informational states* and the resulting state as an *answer* and reformulate the response condition to encompass both types of response. Nothing substantive turns on the terminology.

The state produced by the map-mediated transition is a content-faithful response to the question that is the content of the preceding state, not a merely causal transition of the sort the response condition is meant to exclude. In trying to determine how to get from A to B, I actively search for information of the form *this route gets me from A to B* that will resolve the question. The resulting state, having exactly this content, resolves the question, ends the search, and is thus a content-faithful response to the questioning attitude that initiated the search.

One advantage of this account is that it avoids overintellectualizing the abilities involved. Infants and many nonhuman animals have a hippocampal–entorhinal system capable of constructing cognitive maps and display curiosity behaviors consistent with questioning attitudes. The account does not depend on conceptual, metacognitive, or linguistic capacities that these agents are unlikely to possess. It can therefore illuminate what is rational or intelligent about inquisitive behavior across a wide range of cases.

More mature humans, equipped with concepts such as popularity or competence, can construct cognitive maps along dimensions that extend beyond the purely spatial. Although it remains an open question which dimensions we use in map construction, we can imagine cases—beyond social hierarchies—where abstract dimensions could play a useful inferential role. Take the example of planning a weekend: two salient dimensions of candidate activities are duration and enjoyability. Mapping activities along these axes provides an efficient strategy

for deciding how to spend one's time. Such a map highlights the most relevant metrics, omits extraneous details, and allows for quick filtering of options according to time constraints.

For example, reading a book or cooking a new recipe may offer high enjoyment at moderate duration, whereas a day trip might be even more enjoyable but consumes most of the day. A cognitive map of activities along the dimensions of duration and enjoyment would make it possible to quickly identify options that fit time constraints. For example, commitments to clean the apartment and restock the fridge rule out activities that last the entire weekend. The search can then focus on activities below a chosen time threshold, highlighting those expected to provide the greatest enjoyment within that limit.

4.2. Mental simulation

One key way cognitive maps mediate content-specific transitions is by structuring mental simulations. Replay sequences over cognitive maps coincide with activations of semantic and episodic memories (Higgins et al., 2021; Kaefer et al., 2022), which in turn drive simulations of possible states of the world, often involving sequences of events, actions, or experiences. Mental simulations integrate information from a variety of special-purpose systems—sensory, motoric, affective, evaluative, map-like, and others—to construct suppositional scenarios. Through these simulations, we can combine stored information to answer questions such as: Is the simulated event or action an accurate representation of the world? Is it likely to occur? Is it worth pursuing? Could it serve as an effective means to an end? In this way, mental simulations allow us to draw conclusions about what is the case and what actions to take.

Consistent with this, mental simulations are increasingly recognized as a form of reasoning (Munroe, 2021; Myers, 2021; Shea, 2024a) and as a source of justification or knowledge (Aronowitz & Lombrozo, 2020; Williams, 2021; Miyazono & Tooming, 2024; Myers, 2025). They are constrained by prior representations to ensure that what we simulate is both relevant to our practical and epistemic concerns and likely to correspond to the actual structure of the world. Examples of such constrainers might include forward models for sensorimotor control (Langland-Hassan, 2016; Carruthers, 2025), causal probabilistic generative models (Williams, 2020), and intuitive physics or core cognition systems (Battaglia et al., 2013; Miyazono & Tooming, 2024). Cognitive maps, I argue, constitute another important constrainer.

When past events are reconstructed in memory, the reconstruction is primarily organized around spatial cues. Places tend to change little and in predictable ways over short timescales, so events occurring in the same location often share recognizable features—such as landmarks

or stationary objects—due to shared environmental constraints. By contrast, events that occur at the same time but in different locations may share few similarities, so temporal contiguity alone is often less informative about how events unfolded. Consequently, when memory search and reconstruction are guided by current informational needs, spatial cues typically provide the most relevant guidance (Aronowitz & Nadel, 2025). Because cognitive maps integrate and organize spatial information, they serve as a central blueprint for reconstructing events. Moreover, since recombining elements from memory is critical for constructing counterfactual and future scenarios (Schacter et al., 2012), this function likely generalizes: cognitive maps help constrain the construction of suppositional scenarios, ensuring they remain broadly consistent with the actual structure of the world.

Cognitive maps also structure simulations in virtue of their role in action guidance. Many mechanisms that guide bodily actions are similarly engaged during mental simulation of those actions (Hardwick et al., 2018). Since cognitive maps guide action, they are likely recruited in action simulations as well. Recent work suggests that cognitive maps are constructed to guide action by flexibly combining map components to meet task demands. This involves activating neural structures that code for locations of environmental entities the agent expects to encounter—such as cells representing directions and distances to relevant objects, landmarks, or goal locations (Whittington et al., 2022). The map components are integrated with additional structures representing information about the anticipated steps of the task. Each step activates representations relevant to performing the task at that moment, including subgoals, action opportunities, sensory cues, and internal variables. During extended, multi-step actions, elements from cognitive maps and other specialized systems can be accessed, maintained, and flexibly manipulated in working memory to guide ongoing action and adjust representations to the current situation (Buehler, 2022; Badre, 2025; Whittington et al., 2025).

Consider the task of making a cup of coffee, which involves coordinating a series of steps to achieve the overarching goal. The steps include retrieving coffee grounds, filter, milk, and cup; placing the filter in the machine; adding coffee grounds; filling the reservoir; turning on the machine; pouring coffee; and finally adding milk. Long-term memory stores information about how to execute this routine, including relevant sensorimotor and map-like representations for each step. The executive control system monitors ongoing progress and

continuously adjusts representations in working memory as needed—for example, substituting cream if milk is unavailable or serving the coffee black.[3]

Cognitive maps play a key role in guiding action and enable agents to adapt efficiently to new environments with familiar layouts. Upon entering a kitchen, a cognitive map of the environment forms, representing locations of key entities and geometric relations between them: the coffee machine on the counter, filters and coffee grounds in a cabinet, the water source at the sink, and milk in the fridge. Prominent features such as the sink, fridge, and countertop may serve as landmarks. Once these entities are bound to locations on the map, the agent can compute directions and distances to guide action. During task performance, working memory accesses relevant map components and integrates them with other representations whose content relates to the current step—for example, combining representations of the direction and distance to the milk with motor representations of how to move the body and manipulate objects to obtain it.[4]

The same processes are engaged when actions are simulated. For example, when considering the question of how I make my coffee, I am likely to simulate the relevant action sequence to work out the answer. I thereby transition from a questioning attitude to states representing my coffee-making routine, without having to rely on logically structured representations. The transition, mediated by a combination of representational structures including cognitive maps, responds to the content contained in preceding states in a content-faithful manner. It therefore satisfies the response condition.[5]

A special kind of map-like representation—priority maps—further strengthens the claim that map-like representations can mediate inferential transitions. Priority maps assign priority values to locations in a scene, indicating the extent to which each location should be prioritized for orienting attention. Locations are assigned priority-values depending on their salience and relevance to one's goals and values (e.g., being potential locations for a search target). Attention is then allocated to regions with high average priority-values (Buehler, 2025). Like other mechanisms that guide action and attention, priority maps may also be recruited offline to construct simulations and direct internal attention. For example, when simulating my coffee-

[3] Reasoning is often considered a mental action—something we do rather than something that merely happens to us. Although a full discussion is beyond the scope of this paper, guidance by the executive system helps explain why map-mediated reasoning qualifies as an action (Buehler, 2022).

[4] For details on how to combine representations during action guidance in a content-faithful manner, see Mylopoulos & Pacherie (2017) and Shepherd (2021).

[5] Joshua Shepherd (2021) argues that integrating diverse representational structures to guide intelligent action constitutes a form of practical reasoning. Both reasoning about what to do during an ongoing action and pre-action simulation can be construed as attempts to work out answers to questions about what to do, mediated by many of the same representational structures. They might therefore constitute similar forms of reasoning.

making routine, I may construct a priority map of the imagined scene, assigning high values to the locations of the coffee machine, filter, milk, etc. Attention would be oriented to representations of those objects, shaping the unfolding simulation by strengthening the activation of these representations, maintaining and manipulating them in working memory, and engaging motor representations of how to handle the objects and of the expected sensory consequences of doing so. Shifts of attention mediated by priority maps may thus drive mental simulations forward and, by extension, mediate inferential transitions.

4.3. Responsiveness to rational norms

Let us now turn to the other two conditions for reasoning, starting with the rationality condition. As noted, mental simulations are often constrained by various representational structures, including cognitive maps, in ways that make them likely to correspond to the actual structure of the world. This is often considered sufficient to confer rational or justificatory status on processes involving imagination (Aronowitz & Lombrozo, 2020; Williams, 2021; Miyazono & Tooming, 2024; Myers, 2021, 2025).

Another potential source of rationality lies in the affective states evoked by mental simulations (Gilbert & Wilson, 2007). These states represent the value or disvalue of events and actions, informing us whether we have reason to bring them about or avoid them. Events and actions are continuously appraised against these values representations, eliciting pleasant or unpleasant affective states depending on whether the represented value is positive or negative. Value representations are, in turn, calibrated to track the actual value of events and actions: if an event or action repeatedly yields a higher-than-expected reward, its represented value gradually increases; if it yields a lower-than-expected reward or an unexpectedly unpleasant outcome, its represented value decreases. When properly calibrated, affective states will therefore tend to track the actual value of events and actions (Carruthers, 2025). Simulations accompanied by positive affect therefore indicate reason to pursue the simulated scenario, whereas simulations eliciting negative affect indicate reason to avoid it.

Recent work on metacognition highlights additional ways in which map-mediated transitions can be responsive to rational norms. Evidence suggests the mind continuously monitors the quality and costs of different strategies and learns to favor those that tend to yield the best outcomes in context. This includes learning to select heuristics with the most favorable cost–benefit trade-off for a task (Lieder & Griffiths, 2017), adapt planning strategies to the structure of the environment (Callaway et al., 2022), increase model-based control when it

improves accuracy (Kool et al., 2017), and to regulate interactions between automatic and cognitive control processes as skill develops (Pacherie & Mylopoulos, 2020).

The metacognitive signals guiding strategy selection can take various forms. Shea (2024b) proposes that we are motivated to draw an inference when it is accompanied by a feeling of reliability, whereas we are dissuaded by feelings of unreliability. These feelings are calibrated by their downstream consequences: inferences that consistently yield true conclusions strengthen feelings of reliability—and vice versa for those leading to false conclusions. This mechanism applies to both content-specific and content-general transitions and likely extends to map-mediated transitions.

Similarly, Joshua Shepherd (2025) proposes that metacognitive processes monitor the quality and costs of actions, including mental actions such as reasoning. Quality is represented as the likelihood of success or failure, eliciting feelings of fluency, control, reliability, or confidence. Costs include energy expenditure, working memory demands, time on task, and whether potential benefits justify the required effort or impose excessive opportunity costs. Representations of cost elicit feelings of effort, difficulty, strain, or concentration. Actions expected to offer the best trade-off between quality and cost are most likely to be chosen.

By tracking the value, reliability, quality, and costs of candidate strategies and selecting those that achieve the best trade-offs, these metacognitive processes are responsive to certain rational norms: they optimize for reliability and expected value. Evidence suggests that similar processes govern the deployment of cognitive maps. Liu et al. (2021) found that replay sequences are prioritized according to their expected value: sequences expected to most improve future decisions by updating values of imminent choices are most likely to be replayed. Wimmer et al. (2023) found that, prior to a decision, replay of possible future paths increased when planning held greater benefits. By contrast, after choice feedback was received when no immediate future actions were required, a memory-preservation effect was observed, marked by enhanced replay of paths visited less frequently in the recent past. Many tasks, some involving reasoning, are optimally solved using cognitive maps. It is therefore plausible that metacognitive regulation of how to use cognitive maps extends to reasoning.

Metacognition may also illuminate the notion of taking. Metacognitive feelings are conscious experiences that reflect the quality and costs of candidate strategies and guide one's reasoning accordingly. An agent whose choice of reasoning strategy is accompanied by metacognitive feelings of fluency, control, reliability, or confidence can plausibly be said to take the conclusion to be supported by the preceding states and operations. They appreciate that the conclusion is supported by the states that preceded it, basing their conclusion on these

states because of this fact. Hence, metacognitive feelings of the sort that plausibly accompany map-mediated transitions satisfy the taking condition.

5. Conclusion

We have seen that content-specific transitions mediated by cognitive maps can meet commonly accepted conditions for reasoning: conclusions are responses to premise-states; transitions are responsive to rational norms; and the reasoner can appreciate that the states and operations on which they base their conclusion support their conclusion. Since such transitions are not rule-governed operations over propositional attitudes, this challenges the view that reasoning exclusively takes this form. Instead, we should adopt a more pluralistic view that acknowledges a broader range of representational structures and operations in reasoning.

References

Aguilera, M. (2021). Heterogeneous inferences with maps. *Synthese, 199*(1-2):3805-3824

Aguilera, M. (2025). Journey planning: a cartography of practical reasoning. *Philosophical Explorations, 28*(2):142-164.

Aronowitz, S. & Lombrozo, T. (2020). Learning Through Simulation. *Philosophers' Imprint,* 20.

Aronowitz, S. & Nadel, L. (2025). Space, and not Time, Provides the Basic Structure of Memory. In Lynn Nadel & Sara Aronowitz (eds.), *Space, Time, and Memory* (pp. 95-110). Oxford University Press.

Badre D. (2025). Cognitive Control. *Annual Review of Psychology*, *76*(1):167–195.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45):18327–18332.

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415).

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1):1-18.

Broome, J. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

Buckner, C. (2019). Rational Inference: The Lowest Bounds. *Philosophy and Phenomenological Research*, *98*(3):697-724.

Buehler, D. (2022). Agentive capacities: a capacity to guide. *Philosophical Studies, 179*(1):21-47.

Buehler, D. (2025). The Priority Map. *Australasian Journal of Philosophy*, *103*(1):235–260.

Burge, T. (2018) Iconic Representation: Maps, Pictures, and Perception. In Shyam Wuppuluri & Francisco Antonio Doria (eds), *The Map and the Territory* (pp. 79–100). Springer.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour, 6*(8):1112–1125.

Carruthers, P. (2018). Basic questions. *Mind & Language, 33*(2):130-147.

Carruthers, P. (2025). *Explaining our Actions: A Critique of Common-Sense Theorizing*. Cambridge University Press.

Camp, E. (2007). Thinking with Maps. *Philosophical Perspectives, 21*:145–182

Camp, E. (2018). Why Maps are Not Propositional. In Alex Grzankowski & Michelle Montague (eds.), *Non-Propositional Intentionality* (pp. 19-45). Oxford University Press.

Fodor, J. A. (1987). Why There Still Has to Be a Language of Thought. In *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (pp. 135–154). MIT Press.

Fodor, F. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1-2):3-71.

Frankland, S. M., & Greene, J. D. (2020). Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology, 71,* 273–303.

Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives, 27*(1):145-174.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science, 317*(5843):1351–1354.

Hardwick, R. M., Caspers, S., Eickhoff, S. B., & Swinnen, S. P. (2018). Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience and Biobehavioral Reviews*, *94*, 31–44.

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., & Woolrich, M. (2021). Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron*, *109*(5):882–893.

Kaefer, K., Stella, F., McNaughton, B. L., & Battaglia, F. P. (2022). Replay, the default mode network and the cascaded memory systems model. *Nature Reviews Neuroscience*, *23*(10), 628–640.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, *28*(9):1321–1333.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*:454–469.

Langland-Hassan, P. (2016). On Choosing What to Imagine. In Amy Kind & Peter Kung (eds.), *Knowledge Through Imagination* (pp. 61-84). Oxford University Press.

Levy, Y. (2024). Who is a reasoner? *Inquiry*, 1–27.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124*(6):762–794.

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*(3):640–652.

Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544).

Miyazono, K. & Tooming, U. (2024). Imagination as a generative source of justification. *Noûs*, *58*(2):386-408.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*:e32548.

Munroe, W. (2021). Reasoning, rationality, and representation. *Synthese, 198*(9):8323-8345.

Myers, J. (2021). Reasoning with Imagination. In Amy Kind & Christopher Badura (eds.), *Epistemic Uses of Imagination* (pp. 103-121). Routledge.

Myers, J. (2025). How Imagination Informs. *Philosophical Quarterly, 75*(1):167-189.

Mylopoulos, M. & Pacherie, E. (2017). Intentions and Motor Representations: the Interface Challenge. *Review of Philosophy and Psychology, 8*(2):317-336.

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology, 28*(1):R37–R50.

Pacherie, E. & Mylopoulos, M. (2020). Beyond Automaticity: The Psychological Complexity of Skill. *Topoi, 40*(3):649-662.

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. *Neuron*, *107*(6):1226–1238.

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, *24*(9):1292–1301.

Quilty-Dunn, J. & Mandelbaum, E. (2018). Inferential Transitions. *Australasian Journal of Philosophy*, *96*(3):532-547.

Shea, N. (2024a). *Concepts at the Interface*. Oxford University Press.

Shea, N. (2024b). Metacognition of Inferential Transitions. *The Journal of Philosophy*, *121*(11):597-627.

Siegel, S. (2019). Inference Without Reckoning. In M. B. Jackson & B. Jackson (eds.),

Reasoning: New Essays on Theoretical and Practical Thinking (pp. 15-31). Oxford University Press.

Rescorla, M. (2009) Cognitive Maps and the Language of Thought. *British Journal for the Philosophy of Science, 60*:377–407.

Rescorla, M. (2017). 'Maps in the Head?'. In Kristin Andrews & Jacob Beck, (eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 34–45). Routledge.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, *76*(4):677–694.

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018). Mind-Wandering as a Natural Kind: A Family-Resemblances View. *Trends in Cognitive Sciences*, *22*(6):479–490.

Shepherd, J. (2021). Intelligent action guidance and the use of mixed representational formats. *Synthese, 198*(Suppl. 17):4143-4162.

Shepherd, J. (2025). Salvaging the "sense of agency": Metacognitive feelings for flexible behavioral control. *The Journal of Philosophy*.

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5):1249–1263.

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). How to build a cognitive map. *Nature neuroscience*, *25*(10):1257–1272.

Whittington, J. C. R., Dorrell, W., Behrens, T. E. J., Ganguli, S., & El-Gaby, M. (2025). A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. *Neuron*, *113*(2):321–333.

Williams, D. (2021). Imaginative Constraints and Generative Models. *Australasian Journal of Philosophy, 99*(1):68-82.

Williams, I. (2024). Breaking the language barrier: conceptual representation without a language-like format. *British Journal for the Philosophy of Science*.

Wimmer, G. E., Liu, Y., McNamee, D. C., & Dolan, R. J. (2023). Distinct replay signatures for prospective decision-making and memory preservation. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), e2205211120.

ARTICLE 4

Predictive Minds Can Be Humean Minds

Abstract: The predictive processing literature contains at least two different versions of the framework with different theoretical resources at their disposal. One version appeals to so-called optimistic priors to explain agents' motivation to act (call this optimistic predictive processing). A more recent version appeals to expected free energy minimization to explain how agents can decide between different action policies (call this preference predictive processing). The difference between the two versions has not been properly appreciated, and they are not sufficiently separated in the literature. They constitute two different theories with strikingly different accounts of motivation and action. By reducing all desire-like constructs to belief-like constructs, optimistic predictive processing entails a substantial revision of standard accounts of motivation and action in philosophy and cognitive science. By contrast, preference predictive processing introduces desire-like constructs that play Humean motivational roles in the explanation of action. In this Humean iteration, predictive processing resembles other prominent computational frameworks implementing a distinction between beliefs and desires, such as reinforcement learning and Bayesian decision theory. Ultimately, predictive processing faces a dilemma between parsimony of mental constructs and completeness of its explanations of agency and the mind.

1. Introduction

Over the last decade, predictive processing (PP) theories have enjoyed rapidly growing interest in both philosophy and cognitive science. According to these theories, the agent navigates its environment by engaging in a form of model-based inference aimed at minimizing prediction error. PP's two main selling points are its universality—its potential to provide a unified and mechanistically plausible theory of all neurocognitive phenomena—and its parsimony—its potential to explain all neurocognitive and behavioral phenomena in terms of the same fundamental process: precision-weighted prediction error minimization, or Bayesian inference (Friston, 2009, 2010).

A central appeal of PP to some theorists is its radical conception of agency. Many contemporary philosophical theories of agency are Humean in nature, presupposing a basic distinction between beliefs and desires. Beliefs tell you how the world is, while desires

motivate actions. If PP delivers on its promise, then it seems that an account of agency can be given without reference to desires. All that is needed is a web of probabilistic beliefs that guides perception and action. In other words, PP is anti-Humean.

The anti-Humean commitments of PP, combined with its universalist aspirations, have been the source of considerable debate recently (Klein, 2018, 2020; Clark, 2020; Sun & Firestone, 2020; Van de Cruys et al., 2020; Yon et al., 2020). Critics have pointed out that because of its anti-Humeanism, PP cannot be a universal account of the mind: at some point you need desires to explain motivation and action. The main intuition pump to motivate this challenge is the dark room problem: why does a prediction error minimizing agent not simply seek out highly predictable environments, such as a dark room? In such a highly predictable environment, the PP agent would continuously predict pitch darkness that would always match the sensory input perfectly (Friston et al., 2012; Klein, 2018; Sun & Firestone 2020). Clearly, we are not such agents. The challenge for PP is to explain why we are not dark-room-seeking creatures and why we are instead motivated to go out and explore the world. Such motivation is standardly understood in terms of the agent having mental states, such as desires, with a world-to-mind direction of fit.

Over the years, PP theorists have repeatedly responded that PP has the theoretical resources to handle the dark room problem, as well as other challenges that might seem to require invoking distinct desires. Traditional responses attribute stubborn optimistic prior expectations to the agent that it occupies states that satisfy its bodily needs and curiosity about novel information (Bruineberg et al., 2018a; Yon et al., 2019; Van de Cruys et al., 2020), while more recent responses appeal to more sophisticated models of policy selection via expected free energy minimization (Clark, 2020; Seth et al., 2020). The implicit assumption among both proponents and critics is that PP remains anti-Humean: if PP turns out to be true, we would have succeeded in providing a universal account of agency exclusively in terms of probabilistic beliefs and expectations.

In this paper, we take issue with the common assumption that all versions of PP are incompatible with Humeanism and standard philosophical theories of motivation and action more generally. We argue that PP has recently developed a Humean branch. We pinpoint the origin of this branch to the introduction of models that make use of so-called expected free energy minimization (Friston et al., 2015). Expected free energy models involve both state-estimation, the estimation of the most likely state of the environment given current sensory input, and policy selection, which is the selection of a policy that is expected to lead to preferred

outcomes. We argue that an expected free energy minimizing creature can believe it is in one state while desiring to be in another. Hence, this is a Humean creature.

We identify two distinct branches in the PP literature, the difference between which has not been properly appreciated. According to one branch, which we will refer to as optimistic PP, agents are equipped with optimistic priors that make them predict that they will observe outcomes that are favorable to them. Optimistic predictions and prediction error minimization drive actions toward such outcomes. According to the other, which we will refer to as preference PP, agents score action policies on how well they minimize expected free energy and select the ones that strike the best balance between reducing uncertainty and bringing about preferred outcomes. The theories entail strikingly different accounts of motivation and action. We will conclude that while optimistic PP is anti-Humean and limited in explanatory scope, preference PP has wider explanatory scope, but also comes with straightforwardly Humean commitments, specifically, the acceptance of desire-like constructs. Consequently, the PP theorist needs to choose between parsimony (there are only belief-like states) and universality (explanation extends to all aspects of agency and mental function).

To make the discussion more approachable for non-experts in the PP literature or action theory, we have laid out our argument as follows. In section 2, we present Humeanism, as well as the main challenges for anti-Humean PP accounts of motivation and action. In section 3, we introduce the distinction between optimistic PP and preference PP. In section 4, we show that preference PP is a Humean theory. In section 5, we articulate their respective implications for a theory of action. If optimistic PP and preference PP entail different commitments with respect to the Humean nature of motivation, they imply different theories of action. We argue that while preference PP is compatible with certain versions of standard accounts of action, optimistic PP entails a radical revision. In section 6, we discuss the assertion that adopting the free energy principle (FEP) eliminates all distinctly motivational constructs from our ontology. That is, one might object that preference PP is, in fact, firmly anti-Humean. We argue that the FEP entails no such elimination. In the end, PP is confronted with a choice between universality and parsimony. In its ambition to explain all aspects of agency and mental function, PP has had to invoke desire-like constructs playing Humean motivational roles in the explanation of action.

2. Predictive Processing and Motivation

According to the Humean theory of motivation (Davidson, 1963; Smith, 1987), beliefs and desires are distinct types of mental states. Desires have motivational force, whereas beliefs

have none. We follow Davidson (1963) in understanding desires as standing for the broader category of pro-attitudes. Pro-attitudes in the literature on practical reasoning, action, and ethics are attitudes in favor of something (for example, approval, admiration, liking, preference, and esteem) and include evaluative judgments that an action has some positive characteristic (for example, being desirable, reasonable, admirable, or dutiful). The notion of pro-attitudes thereby goes beyond a simple form of Humeanism restricted to primitive motivational states, such as drives or urges. This becomes important later when discussing the notion of preferred outcomes found in preference PP models, which could include evaluative judgments that play a motivational role and cannot be reduced to a purely doxastic register. Though used interchangeably, we will primarily use the more common term *desires*, except where the broader connotations of pro-attitudes become important.

According to the Humean, reason on its own is never sufficient to motivate agents to act. This idea is often understood as the claim that beliefs on their own are motivationally inert. Only with the addition of some desire will the agent be motivated to act. This difference between motivationally inert beliefs and motivationally active desires is often captured by defining desires in terms of dispositions to act. What we call the Humean theory of motivation is not committed to any particular theory of desire (Schroeder, 2004) or any strict form of motivational externalism (Williams, 1979). The Humean theory is simply committed to the claims that explanation of instrumental action requires both beliefs and desires, that desires are primitive or irreducible to beliefs, and that only desires have motivational force (which can be understood in terms of dispositions to act or in some other way).

One common way of describing the difference between beliefs and desires is in terms of a difference in direction of fit (Searle, 1983). Beliefs have a mind-to-world direction of fit: in case of a mismatch, they ought to be revised to fit the world. Desires, on the other hand, have a world-to-mind direction of fit: in case of a mismatch, the world should be changed to fit the desire. Given Humeanism, if there is a mismatch between the desired situation and the state of the world, the agent should be motivated to act to change the state of the world. Anti-Humeans, on the other hand, reject the existence of distinct desires and their unique role in motivating action. It is often claimed that PP theories are anti-Humean in this sense.

2.1. Humean challenges to predictive processing

According to PP, in all its guises, the agent navigates its environment by engaging in a form of model-based inference (Friston, 2009, 2010; Clark, 2013, 2016; Hohwy, 2013, 2016; Parr et

al., 2022). The brain generates a stream of top-down predictions about what sensory signals it expects to receive given its current best model of the world and the situation in which the agent finds herself. The predicted input is compared to the input the agent actually receives. This comparison leads to prediction errors, which are used to update the generative model or to change the sensory input by acting on the world. The combined process is sometimes called active inference. Since the theory's only primitive is prediction, many have pointed out that it cannot sustain a belief–desire distinction. The lack of distinct desires makes PP anti-Humean.

The fact that an anti-Humean theory of motivation denies the existence of distinct motivational, action-disposing mental states does not entail a denial of the distinction between directions of fit. The anti-Humean could still accept that mental states are sometimes made to fit the world, and sometimes the world is made to fit our mental states. Anti-Humeans often argue that beliefs can have motivational power, which is to say, they can have a world-to-mind direction of fit. Thus, not only do anti-Humeans need to provide reasons for denying the Humean distinction between beliefs and desires, but they also need to offer an alternative explanation of how agents resolve various types of conflict between mental states and the world without resorting to a primitive notion of desire or pro-attitude. Several Humean objections to PP have highlighted this explanatory challenge.

First, the need for action-disposing mental states with a distinctly world-to-mind direction of fit seems to be driving the dark room problem: without a mental state that disposes the agent to change the world and leave the dark room, how can the PP theorist explain the agent's ability to leave the dark room (Klein, 2018; Sun & Firestone, 2020)? This is one way to motivate the Humean challenge to PP. Without distinct belief-like and desire-like states, core aspects of cognition and behavior seem to be left unexplained (Yon et al., 2020).

Second, predictions alone seem to be motivationally inert, and so PP does not seem to have the necessary primitives to explain what motivates action, or how and why we select the actions that we do (Klein, 2018). Frequently, agents must choose between multiple actions and rank them against each other. If the agent does not have distinct motivational states representing the value or reward of different actions, such selection problems quickly become intractable. Representing actions in terms of both probabilities and value allows for simpler comparison than if each action must be specified purely in terms of long series of conditional predictions (*when conditions C_1 and C_2 and C_3, … obtain, I will perform actions A_1 or A_2 or A_3 …*).

Third, other computational frameworks, including reinforcement learning and Bayesian decision theory, are consistent with Humeanism. In these frameworks, value signals are always necessary for action, and probabilities and values are represented and computed independently.

The empirical success of these frameworks provides some empirical support for Humeanism (Colombo, 2017).

Finally, a fourth challenge to anti-Humean PP has recently been presented by Klein (2020). According to Lewis (1988, 1996), desires are contingent. We might have uncommon desires or lack common ones. There is no necessary relationship between desire and belief. As Lewis (1996, p. 304) puts it: 'Any values can go with any credences'. Building on this insight from Lewis, Klein (2020) has argued that PP fails to simultaneously explain action and value learning. To explain action under PP, the predictions driving actions must be effectively unrevisable to ensure that the ensuing prediction error can only be minimized through action instead of by simply revising the predictions.[1] However, sometimes action-guiding predictions ought to be revised when the agent learns that an action is no longer desirable. For example, when the agent learns that the water is contaminated, she ought to revise the prediction that *when I am thirsty, I drink water*. In short, to explain action within PP, the predictions driving action need to be unrevisable—but to explain value learning, they need to be revisable. This problem could be resolved by adding a rule that allows the agent to update the prediction when the water is observed to be contaminated.

However, to ensure that the prediction remains unrevisable in normal circumstances (where the water is not contaminated), the updating rule cannot be Bayesian updating. If the prediction were generally susceptible to Bayesian updating, the agent could update it, even in normal circumstances. For example, the agent could take prolonged periods of thirst without drinking as evidence against the prediction that *when I am thirsty, I drink water*. Beliefs, therefore, need to include lots of conditionals so that the updating rule only applies in the right circumstances (for example, *unless the water is contaminated, when I am thirsty, I will drink water*). But such conditionals can get extremely complex, even for relatively simple creatures. They require specifying a virtually infinite number of conditions for when beliefs should and should not be updated, for every scenario the organism might find itself in. It is highly implausible that our beliefs have this kind of complexity. Moreover, it rests on the dubious assumption that we come pre-wired with complex expectations fit for nearly every possible scenario, instead of adapting to environmental changes by learning about changes in value. Klein (2020) points out that Humean theories have an easier time explaining such cases. When we learn that the water is contaminated, we can leave other beliefs untouched and simply update our desires (for

[1] Unrevisability can be achieved by taking the prediction to be so precise that no amount of prediction error will suffice to revise it. For an account of motivation in terms of unrevisable predictions, see Miller Tate, (2021).

example, *store-bought beverages are now strongly preferable to tap water*). Beliefs can thus remain responsive to the evidence via Bayesian updating, while desire updates ensure that we can still adapt and pursue the best course of action when circumstances change.

2.2.The revisionist response

Some proponents of PP have argued that PP can in fact explain the phenomena targeted by Humean critics without the need to posit distinct motivational states. Clark (2020) argues that motivational states can be cast as counterfactual predictions, the content of which is what we would observe if we acted in a certain way. We counterfactually predict that we are already in the desired state. This initially gives rise to prediction error, which is minimized by bringing the agent into the desired state. According to Clark (2020, p. 12), PP should treat desires as 'varying forms and time-scales of prediction', the motivational force of which is dictated by the relative precision-weighting of those predictions. These counterfactual predictions simultaneously have belief-like features (they predict what will occur as a consequence of the action) and desire-like features (they are poised to bring about the predicted consequences of acting). This interpretation of PP essentially reduces all mental state types to a single primitive: precision-weighted predictions.

To summarize, critics have raised a number of challenges to PP based on its anti-Humean commitments, while its proponents have argued that PP can meet these challenges by revising our Humean intuitions. What matters for our purposes is the shared assumption that PP is inescapably anti-Humean. The assumption that PP does not allow for distinct types of mental states is not just held by philosophers working on PP, but is repeated in the more technical literature (Friston et al., 2009; Friston et al., 2012; Friston et al., 2015; Friston, 2019; Parr et al., 2022). Specifically, it is sometimes claimed that PP entails a kind of desert landscape ontology, where desires, goals, reward signals, and the like do not exist and are replaced by more parsimonious models of purely prediction-driven embodied exchanges of creatures with their environment (Friston, 2019).

3. Predictive Processing, the Devil, and the Details

Thus far, we have discussed PP in broad outline as a framework according to which everything is precision-weighted prediction error minimization. We would now like to argue that the current discussions have failed to recognize that PP architectures come in different guises,

which account for motivation and action in different ways. To make this more explicit, we introduce a distinction between two theories: optimistic PP and preference PP.

3.1.Optimistic predictive processing

PP accounts vary in what they take the main purpose of the generative model to be. Initially, PP accounts were seen as a continuation of a Helmholtzian view of perception (Friston et al., 2012; Clark, 2013; Hohwy, 2013). On such an account, the main purpose of PP is to reconstruct the hidden state of the world based on proximal sensory input. The content of perception is then determined by the set of predictions that manage to explain the sensory signals, or, equivalently, that manage to explain away the prediction errors. Following the Helmholtzian approach, action is seen as a kind of experiment that disambiguates between competing hypotheses and increases the evidence for one's current hypothesis (Friston et al., 2012). As illustrated by the dark room problem and the other Humean challenges discussed in the previous section, it is difficult to see how a purely prediction-driven version of PP can explain all aspects of agency and deliver a unified theory of the mind.

One way to respond is to argue that agents are endowed with so-called optimistic priors, which dispose the agent to predict favorable outcomes, consistent with having their bodily needs met (for example, a full stomach, stable blood-glucose and hydration levels, and a body temperature of around 37°C). These priors are 'optimistic' in that the agent expects to observe outcomes that are 'good' for the agent, in the sense that they are compatible with its continued existence. They are also sometimes referred to as stubborn predictions (Yon et al., 2019), as they resist revision and can only be satisfied by making the agent's observations conform to its expectations. There is a second sense in which the agent's priors are 'optimistic': the agent expects a beneficial environment in which all its bodily needs can be met. Typical environments are not like this; they might be cold or lack food and water. To act adaptively, the stubborn agent needs to keep believing that it will find beneficial environments. Hence, adaptive action involves what Wiese (2017) calls 'systematic misrepresentations of the environment': the agent's expectations are (and need to be) systematically skewed toward the kinds of environments in which it thrives.

Some simulation studies of PP principles, where agents learn the structure of an environment, simply assume that the environment is beneficial. In Friston et al. (2009), for example, the agent is first shown the correct sequence of actions without being able to intervene. After having learned the optimal solution in a controlled environment, the agent is

then endowed with the capacity to act, and to make its observations congruent with the previously learned sequence of observations. The idea is that an agent equipped with optimistic priors will not discover the causal regularities in its environment, and update its model accordingly, but will instead make the environment conform to its expectations (Bruineberg et al., 2018a; Yon et al., 2019). To get out of the dark room, a PP agent needs to stubbornly predict that the world is different from how it is currently observed to be. If the agent is equipped with optimistic and stubborn expectations of a full stomach, then, as the agent in the dark room grows hungry, this leads to prediction errors that can only be minimized by leaving the room to eat. Thus, optimistic priors serve as self-fulfilling prophecies that compel the agent into action, even if this means facing less predictable environments. Since the defining construct of this version of PP is optimistic priors, we will refer to it as *optimistic PP*.[2]

Optimistic PP runs into the problems raised by Humean critiques. An account that needs priors to be both unrevisable and systematically skewed to account for adaptive action will have trouble providing an empirically adequate account of value learning: to learn values, the agent's priors need to be revisable. There is, however, a different version of PP on the market with a different architecture.

3.2.Preference predictive processing

To our knowledge, Friston et al. (2015) provide the first articulation of a PP theory that involves the minimization of prediction error in the future; that is, minimization of expected free energy. At first glance, the introduction of expected free energy involves more of the same: 'Our basic approach is to cast optimal behavior in terms of inference, where actions are selected from posterior beliefs about behavior. This allows one to frame goals and preferences in terms of prior beliefs, such that goals are subsequently fulfilled by action' (Friston et al., 2015, p. 188). The authors seem to argue for an anti-Humean position: pro-attitudes, including goals and preferences, reduce to doxastic states, particularly prior beliefs. However, the devil is in the details. To see this, let us unpack the commitments of expected free energy minimization.

An agent has a finite number of policies or strategies available. To rank the policies, the expected free energy of each policy is evaluated. Heuristically, it amounts to the evaluation of the following counterfactual: *What is the free energy I expect to receive if I were to pursue this*

[2] In some discussions of active inference, the explanation of optimistic priors is delegated to the FEP. The FEP holds that any system that maintains its organization over time will engage (or appear to engage) in a form of model-based inference in which the generative model embodies the optimal state of being for that system. We discuss the implications of the FEP in section 6.

policy? Calculating the result for each policy gives a policy-specific expected free energy. The probability of pursuing a policy is then proportional to the relative expected free energy of the policies: probability of policy ∝ expected free energy of policy. In other words: *I will pursue those policies most often that I expect will minimize free energy*. But what exactly is expected free energy? One way to decompose expected free energy is as follows: expected free energy = expected ambiguity + risk.

The expected ambiguity term roughly captures: *How much uncertainty will be reduced by pursuing this policy?* A policy that brings the agent to a location where the agent expects it can gain new information (reducing uncertainty) will have lower expected ambiguity than a policy that brings the agent to a location where the agent does not expect to learn anything new. The risk term roughly captures: *How close will following a particular policy bring me to a preferred outcome?* Here, a lot of the explanatory work is done by the notion of preferred outcomes. So, what is a preferred outcome? In their introduction of expected free energy minimization, Friston et al. (2015, p. 188) define preferred outcomes as follows: 'In active inference, constructs like reward, utility, epistemic value, etc. are described in terms of prior beliefs or preferences. In other words, preferred outcomes are simply outcomes one expects, a priori, to be realized through behavior (e.g., arriving at one's destination or maintaining physiological states within some homoeostatic range)'.

A preferred outcome is thus an outcome the agent expects given the kind of agent it is. The agent is set up to bring about expected outcomes by selecting policies that it expects will lead to those outcomes. If the agent's preferred outcome is *tasting coffee*, then a policy that involves pouring oneself a cup of coffee will involve less risk than a policy that doesn't. Preferred outcomes are defined in terms of a probability distribution over observations some time into the future. For this reason, they are also frequently referred to as *preferred observations*. The two terms are used interchangeably in the literature and refer to the same formal construct: a probability distribution over observations at some time point in the future. For simplicity, we will stick to the term *preferred outcomes*.[3]

If you think preferred outcomes sound suspiciously Humean, you would be correct. In introducing the notion of preferred outcomes, Friston et al. (2015) contrast it with pro-attitudinal constructs like reward and utility, but also likens it to pro-attitudes like preferences.

[3] Other terms used include *preferences*, *preferred states*, and *preferred sensations*. We take it that these are all used interchangeably.

We will return to this question shortly, but for now, let us point out some advantages of the version of PP that operates with expected free energy minimization.

First, minimizing expected free energy involves selecting policies that are expected to lead to new information and preferred outcomes. Another way to put this is that an agent trying to minimize expected free energy is trying to strike an optimal balance between explorative behaviors (reducing expected ambiguity) and exploitative behaviors (reducing risk). In the absence of ambiguity, the agent will simply select the policy that leads to preferred outcomes. In the absence of preferred outcomes, the agent will select the policy that reduces most uncertainty. Taken together, the minimization of expected free energy provides a relatively simple account of action selection in uncertain environments.

A second major advantage of these models is that they add a capacity for planning and decision-making. Expected free energy allows the agent to score different policies about how to act in the future. It allows the agent to consider possible future observations, as well as how possible future observations are conditioned on the policies that the agent could pursue. It adds an internal loop that considers possible future observations and evaluates them relative to preferred outcomes. The inferred policies that strike the optimal balance between bringing the agent toward preferred outcomes and reducing uncertainty will be considered most probable, and, therefore, be selected for action. Since the defining construct of this version of PP is preferred outcomes (at least in the context of explaining motivation), we will refer to it as *preference PP*.

4. Preference Predictive Processing: Going Humean

Let us return to our main question: did PP grow a Humean branch with preference PP? Consider a chess player who selects a move because its consequences are close to the kinds of consequences the player would like to see from a move (that is, seeing her opponent's position crumble rather than her own). The comparison of expected consequences of following an action with preferred consequences is an indispensable element of policy selection using expected free energy minimization. If there is a Humean, desire-like, pro-attitudinal element in preference PP, it is to be found in preferred outcomes.

4.1. Preferred outcomes

How are we to understand the notion of preferred outcomes? Can desires be replaced by beliefs about observations? Opinions on this seem to be mixed. In a recent treatment, Parr et al. (2022) offer the following proposal:

> […] using the notion of expected free energy amounts to endowing the agent with an implicit prior belief that it will realize its preferences. Hence, *the agent's preference for a course of action becomes simply a belief about what it expects to do, and to encounter, in the future*—or a belief about future trajectories of states that it will visit. This *replaces the notion of value with the notion of (prior) belief*. This is an apparently strange move, if one has a background in reinforcement learning (where value and belief are separated) or Bayesian statistics (where belief does not entail any value). (p. 53; emphasis added).

These remarks suggest that talk of *preference* and *value* is somewhat deceptive or derivative. If preferences are fundamentally just beliefs about what the agent expects to do, then preference PP does not seem able to draw a genuine distinction between beliefs and desires, in anything but name. If so, preference PP seems, like optimistic PP, to be firmly anti-Humean.

However, one should be careful not to take such discourse at face value. Generally speaking, the term *belief* in PP does not stand for any standard folk-psychological concept. Instead, a belief, in the technical sense employed in PP and Bayesian inference, is a probability distribution over a set of states. Importantly, the Bayesian notion of belief does not entail the mind-to-world direction of fit.[4] Indeed, this could not be the case, because the direction of fit between the world and Bayesian beliefs depends on the relative precision of predictions and observations. This means that whether a Bayesian belief has a mind-to-world or a world-to-mind direction of fit depends on the decision architecture in which it is embedded.

So, what is the decision architecture in which preferred outcomes are embedded? Here it is worth taking a closer look at the details of expected free energy minimization. The canonical implementations of expected free energy minimization make use of partially observed Markov decision processes (Friston et al., 2015; Friston et al., 2017). Partially observed Markov decision processes make a number of assumptions about the conditional dependencies involved in the decision process. Most notably, they assume that observations at time t (o_t) are only dependent on the current hidden state (s_t), and that the probability of a hidden state s_{t+1} is

[4] This is especially true of Bayesian beliefs over policies, which depend upon preferences over counterfactual outcomes.

dependent only on the previous hidden state s_t and the policy $\pi(t)$. By exploring its environment, an agent can learn the conditional dependencies of its environment: *given that I am in this location, I expect to observe this*, or *given that I am in this location, and plan to walk in this direction, I expect to end up in this other location*. In the terminology of the PP literature, the transition probabilities from hidden states to observations are provided by matrix A, while the transition probabilities between hidden states, conditioned under each policy, are provided by matrix B (Friston et al., 2015, 2017). Preferred outcomes are provided in a separate matrix C.[5] Standard implementations of expected free energy minimization using partially observable Markov decision processes provide the update equations for how the free energy minimizing agent should change its beliefs about the statistical structure of its environment (for a derivation of those equations, see, for example, Bruineberg et al., 2018b, table 2, appendix B).

A few points are worth emphasizing. First, and crucially, the beliefs about the structure of the environment are kept apart from the preferred outcomes that drive the agent's policy selection. The former are stored in matrices A and B, while the latter are stored in matrix C. Second, whereas canonical implementations provide update equations for matrices A and B (that is, equations that capture how the agent updates its beliefs about the environment after observation), far less has been written on implementations of expected free energy minimization that provide update equations for matrix C.[6]

In summary, preference PP can explain motivated behavior through a set of states, specifically, preferred outcomes, with the following properties:

1. *Preferred outcomes are used as a benchmark in policy selection*: How close will following a particular policy bring the agent to its preferred outcomes? The probability of selecting a policy is proportional to its proximity to preferred outcomes.
2. *Preferred outcomes are independent of the beliefs the agent has about the causal and statistical structure of its environment*. To the extent that preferred outcomes have updating rules, these are independent of the rules for updating beliefs about the environment.

[5] Although this labelling of the transition probabilities seems specific to the PP literature, the general idea of a partially observed Markov decision process is well established. A standard machine learning textbook introduces them as 'basically a hidden Markov model augmented with action and reward nodes' (Murphy, 2015, p. 331). Transitions between hidden states are conditioned on the agent's actions, and these actions themselves are selected in accordance with the agent's rewards. In preference PP terminology, transitions are conditioned on policies, and policies are selected in accordance with preferred outcomes.

[6] One exception is the work on preference learning by Sajid et al. (2021). We return to this in section 4.2.

In other words, these mental states guide action selection, have a world-to-mind direction of fit, and are updated independently of the updating of beliefs about the structure of the environment. They thus have all the functional characteristics of desires. By contrast, the beliefs about the structure of the environment have all the functional characteristics of beliefs. Note that reflecting the nature of the creature, preferred outcomes could take the form of a wide range of pro-attitudes from basic drives and urges (for example, for food and shelter) to higher-level evaluative judgments (for example, about the favorability of chess positions or moral actions). However one fills in preferred outcomes, preference PP allows for the existence of distinct desires (broadly construed) that motivate action over and above belief-like states. Hence, preference PP is consistent with Humeanism.

To illustrate the basic points, consider an agent engaging a very simple blue or red environment. Let us assume that the values stored in its *C*-matrix are such that the agent has *perceiving red* as a preferred outcome and *perceiving blue* as an undesirable outcome. The agent starts out not knowing which parts of the environment are blue and which are red. As it explores its environment, it only ever encounters blue. The agent correctly infers that it inhabits a blue environment. This knowledge gets stored as a prior belief in its *A*-matrix (which stores the probabilities of observations given one's current state): *given that I am in this location, I expect to observe blue*. But throughout this exploration, its preferred outcomes (stored in the *C*-matrix) remain unchanged. It still has *perceiving red* as a preferred outcome and *perceiving blue* as an undesirable outcome. The simple fact that the agent can maintain these preferences while learning that its environment only contains blue implies that the agent is a Humean creature. It has distinct beliefs and desires that are updated independently.

4.2. Preference learning

In our discussion so far, we have presupposed that the updating rules for the *C*-matrix will need to be substantially different from the updating rules for the *A* and *B* matrices. After all, the *A* and *B* matrices try to approximate the structure of the agent's environment, while the *C*-matrix captures the agent's preferred outcomes. As detailed above, these are very different functional states. A crucial point is that to benefit from the resources offered by the distinction between belief-like and desire-like states, a Humean agent needs to have independent updating rules for both types of states. Consequently, in line with Klein (2020), value learning needs a non-Bayesian updating rule.

Sajid et al. (2021) recently proposed a framework for preference learning that defines updating rules for the C-matrix, showing how PP agents can learn preferred outcomes to guide policy selection without relying on pre-specified preferences.[7] On this account, the agent learns its preferred outcomes by engaging with its environment in much the same way that it learns about the structure of its environment: preferred outcomes are learned via Bayesian updating. To better understand the behavior of such an agent, let us examine the details of the simulations by Sajid et al. (2021) of this learning process.

When placed into an unfamiliar environment, the agent is initially uncertain about the structure of the environment and its own preferred outcomes. At first, the agent engages in purely exploratory behavior and learns the structure of the environment (that is, what to expect where). Next, the agent is equipped with the ability to learn preferred outcomes, and the outcomes observed more often are the outcomes the agent learns that it prefers. As the agent becomes less uncertain about what its preferences are, it will move away from exploring various outcomes and start to seek out its learned preferences.

What are we to make of such an updating rule for preferences? Let's transpose our PP agent to Italy where it spends considerable time in a village with only one restaurant. Not knowing what it wants, or what the items on the menu mean, the agent randomly picks a different item from the menu every night (careful not to order the same thing twice). This is not a bad strategy: the agent now knows what is on offer and how it tastes. In a separate second phase, the agent starts to learn its preferences. It does so by picking a random item from the menu each evening and by keeping a tally of its choices. Over time, it observes itself choosing spaghetti slightly more often than other dishes and starts to bias its ordering toward spaghetti, leading to even more observations of eating spaghetti. At some point, having eaten spaghetti consistently for several days in a row, the creature comes to the inevitable conclusion (paraphrasing Sajid et al., 2021, p. 30): *I am the sort of creature that enjoys eating spaghetti*, and happily eats its favorite food for the rest of its life.

Such an account of value learning seems deeply unsatisfactory. First, it is implausible that preferences are solely determined by the relative frequencies of outcomes encountered. Might the spaghetti-eater not eventually grow tired of spaghetti? The problem runs deeper: while it is true that we often learn what we like by sampling different options, this does not fully explain preference formation. For example, presented with two items on a menu, we may like one and dislike the other—simply because one tastes good and the other does not. A purely Bayesian

[7] We would like to thank two reviewers for bringing this literature to our attention.

account of value learning cannot capture this distinction. In reality, repeated exposure is not necessary to acquire preferences. Sometimes we immediately learn to prefer or disprefer certain outcomes (we can confidently judge good and bad taste after a single bite), and sometimes preferences are hard-wired (for example, aversion to pain). Bayesian updating over multiple rounds of observations will, therefore, often be the wrong place to look for preferences. Sajid et al. (2021) acknowledge the counterintuitive consequences, or potentially suboptimal strategies, implied by their Bayesian updating framework. As demonstrated by their simulations, in an environment where the agent encounters obstacles more frequently than the goal state, the PP agent will learn to prefer and seek out the obstacles rather than the goal state. Hence, this updating rule risks teaching the agent plainly counterproductive strategies.

On their account, preference learning is essentially a form of Bayesian updating applied to the preferred outcomes stored in the *C*-matrix. Preferred outcomes are cast as prior beliefs that the agent will encounter those outcomes, and the updating rules for likelihood and preferred outcomes are the same (they are both the experience-dependent updating of concentration parameters of a Dirichlet distribution). This suggests that preferred outcomes are not relevantly distinct from the beliefs about statistical regularities in the environment stored in matrices *A* and *B*. Consequently, the anti-Humean's problem resurfaces: we need an explanation of our ability to keep beliefs about statistical regularities fixed while independently revising our desires. The fact that preferred outcomes are stored in a separate matrix does not seem to help by itself. If prior beliefs across matrices are sensitive to the same evidence and are all revised via Bayesian updating, then preferred outcomes and other prior beliefs should be revised in tandem and converge over time. Indeed, the spaghetti-enjoying creature will both like and expect spaghetti every evening. Given the updating rules it is subject to, it seems difficult to explain how it can end up liking one thing and expecting another.

To conclude, a view on which preferred outcomes are updated via Bayesian updating faces a dilemma. If preferred outcomes are simply a form of prior beliefs subject to Bayesian updating, it becomes hard to provide a mechanism for value learning that is independent of general belief updating, and thereby addresses the problems associated with anti-Humeanism. By contrast, if preferred outcomes are not updated in tandem with beliefs about the structure of environment, the agent is somehow able to weigh evidence differently and arbitrate between updating its beliefs and preferred outcomes. Consequently, either PP insists on the parsimony of updating rules and is unable to explain all aspects of agency, or it insists on a broad explanatory scope, which requires independent value learning. In the latter case, we relax the

strict Bayesian updating requirement and allow that preference PP is consistent with a Humean account of motivation.

The discussion above shows that the question of preference PP's Humeanism requires answering two questions. First, is there a desire-like element in preference PP that is separate from belief-like states? We have argued in section 4.1 that preference PP does contain a separate desire-like construct. Second, even if there is such a distinction, are the updating rules for both types of states sufficiently different to make use of that distinction? There are two options here. The first is to follow Sajid et al. (2021) in trying to subsume value learning under Bayesian inference. We have seen that this leads to an account of value learning that must confront the problems associated with anti-Humeanism. The other option is to treat preferred outcomes as a placeholder for some to-be-specified value learning mechanism. With the exception of the purely Bayesian account of preference learning discussed above, the current literature on preference PP models seems to leave open this approach. For PP to avoid the problems facing anti-Humeanism, whatever learning mechanism is slotted in ought to make preferred outcomes independently updatable from beliefs. This will likely require distinct value representations that do not reduce to prior beliefs, along with a non-Bayesian learning mechanism thereof. Both are characteristics of other prominent computational frameworks, such as reinforcement learning and Bayesian decision theory. In fact, nothing seems to prohibit preference PP from adopting an account of value learning from alternative frameworks except the aspiration for a theory of the mind that is both universal and based purely on prior beliefs and Bayesian updating.

5. Two Theories of Action

Pro-attitudes and their motivational role are central to standard accounts of agency. Some version of the Humean theory of motivation is assumed by most theories of action. The difference between optimistic PP and preference PP, and their respective explanatory potential with respect to agency, comes clearly into focus by looking at their implications for a theory of action. In this section, we outline the respective theories of action that optimistic PP and preference PP have on offer. Fundamental to both theories of PP is that all aspects of agency, including motivation and action selection, are cast as inference problems (Friston et al., 2012; Friston et al., 2013; Clark, 2020). Beyond that, the theories have quite different stories to tell.

5.1. Optimistic predictive processing: A revisionist theory of action

In optimistic PP, all mental state types are reduced to a single construct, namely, precision-weighted predictions.[8] To perform an action, the brain predicts that it is currently receiving the sensory input it would expect to receive if the action had already been performed (for example, the proprioceptive signals associated with having raised one's arm). These prediction errors are minimized by activating processes that move the body toward the predicted state. Descending predictions can thus serve as motor commands. They activate motor processes by triggering reflex arcs (neural pathways that control reflexes), which move the body (for example, by contracting muscles in the arm) toward a state where it receives the (proprioceptive) sensory signals associated with the predicted state (Adams et al., 2013; Grünbaum & Christensen 2024). Again, this assumes that the agent's predictions are stubborn in the face of prediction errors, so that the agent is driven to act instead of revising the predictions.

The stubbornness of predictions is accounted for by their precision: when the precision of a prediction is high, it resists revision and causes the agent to pursue the action that makes it come true. Optimistic PP thereby accounts for some aspects of motivation: precise predictions can drive action. If all goes well, the agent is able to prioritize those predictions that help her navigate her environment and meet her needs (Pezzulo et al., 2015, 2018; Clark, 2020). This context-sensitive prioritization of predictions is non-trivial. The challenge is to provide an empirical account without presupposing an agent that can tweak its precision, as it sees fit. One issue with explaining action by a voluntary and context-sensitive tweaking of precision is that it turns a presumably sub-personal mechanism into a capacity governed by the person's will. Not only does this sneak motivation and desire in through the back door, but it also seems to commit the homunculus fallacy of trying to explain agency by positing another agent inside the agent, which itself requires explanation. This puts a heavy explanatory burden on precision.

Assuming that a proponent of optimistic PP can provide a satisfactory anti-Humean account of precision-weighting, such an account of precision needs to stay within the Bayesian commitments of the framework. Given both prior expectations, the prior precision over those expectations, and sensory input, the posterior expectations and precision over those expectations should approximate Bayesian inference. The resulting account would clearly be revisionist, as it compels us to revise standard conceptions of motivation and action. We have already seen how optimistic PP clashes with the Humean theory of motivation, unlike other

[8] Some might prefer the alternative phrasing that the single fundamental construct is the prior beliefs about states and state transitions that generate predictions. We do not think much hinges on the choice of terminology here.

computational frameworks, such as reinforcement learning and Bayesian decision theory. In these frameworks, probabilities and values are represented and computed independently, and value signals are necessary to motivate action.

There are also other aspects of agency that optimistic PP struggles to explain. Consider Buridan cases, where an agent faces a choice between multiple incompatible options that are equally desirable and probable. The canonical case is of a donkey placed right between two equally attractive bales of hay, having to make a choice so as not to starve. Arguably, we often face choices like this in real life (for example, when faced with the choice between equally desirable holiday destinations or items on a menu). The optimistic PP equivalent of such cases seems to be an agent facing two incompatible courses of action, the information about which is equally precise. How can the optimistic PP agent break the tie? It seems puzzling how a predictive brain that deals solely in predictions could come to favor one option when the precision of all options is identical.[9] According to some, such cases require distinct intentions to break the tie and motivate action (Bratman, 1987). To overcome the impasse, we arbitrarily form an intention to pursue one option. The intention then motivates us to act in accordance with it, and structures further planning and practical reasoning. To some, this suggests that intentions cannot be reduced to belief-desire pairs, since such pairs cannot play the same roles in planning and practical reasoning. However, there appears to be no element in optimistic PP that could play the role of intentions to resolve ties in Buridan cases and in structuring further planning and practical reasoning.

Optimistic PP might respond that if the Buridan agent is able to break the tie, it would merely demonstrate that the agent has deep or evolving priors that somehow make the expected precision of one option relatively higher.[10] This effectively amounts to denying the possibility of Buridan cases, because there would always be some prior to the effect that the options are not truly considered equally desirable and probable. Furthermore, this response would appear to settle optimistic PP with some additional problems. This strategy reads off the agent's

[9] Ransom et al. (2017) raise a similar critique. They argue that PP cannot explain our ability to voluntarily shift attention between two overlapping film-streams when the signals from each stream are equally precise.

[10] This is essentially Clark's (2017) response to Ransom et al. (2017): Your voluntary shift of attention takes the form of a counterfactual prediction that you are currently perceiving one of the film-streams, which, in a self-fulfilling manner, increases the expected precision of inputs from that stream, thereby making you perceive that stream. Clark suggests that such a counterfactual prediction can be understood as a desire to see one film-stream rather than the other. Since Clark attributes both belief-like and desire-like roles to a single primitive (precision-weighted predictions), his explanation is of the optimistic PP and anti-Humean variety. The problems facing anti-Humeanism seem just as pressing for mental actions, such as voluntary shifts of attention, as for any other action type. Moreover, as explained in the main text, since each option can be considered equally desirable, distinct intentions are arguably needed in addition to beliefs and desires to break the tie.

preferences from her choices. But if this is how preferences are determined, there seems to be no room for any divergence between motivation and action. Some accounts of action reject the claim that actions are always caused by the strongest desire and argue that agents can act against their strongest desire. That is, agents are able to do something they find truly undesirable or refrain from doing what they find truly desirable (Schueler, 1995). Accepting a distinction between desire and intention enables this type of divergence, because intentions can then play the role of controlling actions, even when they run counter to our strongest desires (Holton 2009). Yet with only precision-weighted predictions at its disposal, this is not a divergence optimistic PP is able to accommodate. In sum, optimistic PP is not incompatible with both Humeanism as well as accounts of action that distinguish intentions from belief-desire pairs.[11]

5.2. Preference predictive processing: A non-revisionist theory of action

According to preference PP, actions are brought about through the inference and selection of optimal policies. Policy selection is the process of inferring which policy minimizes expected free energy; that is, which policy strikes the optimal balance between reducing uncertainty and leading to highly weighted preferred outcomes (Pezzulo et al., 2018; Parr et al., 2022). This process requires a generative model that can model more and more abstract relations between actions and outcomes and score them on how well they minimize expected free energy.

As argued earlier, preference PP is consistent with a distinction between beliefs and desires. To make clearer how preference PP relates to standard conceptions of agency, let us take a closer look at policy selection under preference PP. Expected free energy is composed of both expected ambiguity and risk. Expected ambiguity encodes expectations about how much uncertainty will be reduced by pursuing a certain policy or how much information we stand to gain under a certain policy. Risk encodes expectations about the preferred outcomes a certain policy might bring about.

The evaluation of expected ambiguity is made possible by the fact that an agent has access not just to what it believes, but also to its uncertainty about its beliefs. Expected ambiguity evaluation is therefore dependent on a particular kind of belief: *given that I am in this state and execute this policy, I expect to observe a state with this much uncertainty*. The expected ambiguity term promotes policies that lead to observations that reduce uncertainty, while

[11] A reviewer suggested another response to Buridan cases: it might be that intrinsic noise disturbs the equilibrium. But this simply amounts to denying the phenomenon. If the equilibrium is disturbed by noise, the agent will not be faced with a choice between equivalent options. Our argument is that if Buridan cases exist, they pose a challenge to optimistic PP. We have some reasons for thinking that such cases exist. Hence, claiming that they do not requires a substantial argument.

penalizing policies that do not, and from which the agent, therefore, expects to learn little. The evaluation of the risk is made possible by two different kinds of mental states. On the one hand, there are beliefs of the form *given that I am in this state and execute this policy, I expect to observe this state*. On the other hand, there are desires, or preferred outcomes, of the form *I want to observe this state*. The risk term promotes policies that lead to observations that match preferred outcomes, while penalizing policies that do not.

Some critics might argue that the best interpretation of preferred outcomes is something like: *given that I am this kind of creature, these are the kinds of things I expect to observe* (note that these expectations need not be conscious). Hence, one might argue that preferred outcomes are more akin to beliefs (or predictions) than desires, and therefore, preference PP does not contain distinct desire-like states (we will discuss this objection further in section 6). In section 4, we argued that preferred outcomes are (1) used as a benchmark for action selection, and (2) independent of beliefs the agent has about the causal structure of its environment.

A state with these characteristics is best understood as a desire, not as a kind of belief. The doxastic gloss of preferred outcomes one sometimes encounters in the literature simply mischaracterizes the role preferred outcomes actually play in the preference PP architecture.

By separating out policy selection from state estimation, preference PP has the theoretical tools for both the modulation of precision of sensory signals and the modulation of precision of policies (Parr & Friston 2017, 2019). The distinction between the two forms of precision modulation allows for more flexibility. For example, an agent that has a precise high-level policy of following a diet—that is to say, is very motivated to follow a diet—will lower the expected precision of low-level gustatory outcomes related to consuming high-calorie foods and increase the expected precision of signals and policies related to eating healthier alternatives (Pezzulo et al., 2018). This explains how one action (for example, eating the healthy option) is selected over alternatives (for example, eating cake). Where the optimistic PP agent needs to stubbornly predict that she will not eat the cake, the preference PP agent can infer that eating the cake will meet certain preferences (for example, for high-calorie foods), but nonetheless opt for an alternative diet-congruent policy by making the alternative highly precise, to the point of making it the optimal policy. In short, higher precision of policies translates to higher motivational force.

How about intentions? Recently, some have developed a notion of intentions within a preference PP framework. Friston et al. (2025) argue that in intentional behavior, the agent tries to bring about so-called preferred latent states when selecting policies. Latent states are the presumed but not themselves observable causes of sensory input. A preferred latent state is,

they suggest, simply a prior belief that the agent will bring that state about. In action theory, intentions are often distinguished from beliefs and desires by their distinctive functional and normative roles in practical reasoning and planning. If intentions are simply prior beliefs over latent states, it is not clear that they are sufficiently distinct from beliefs and desires to play the distinctive roles often attributed to intentions.

Another potential interpretation is to identify intentions with selected policies. Since selected policies result from optimal belief-desire pairs (that is, those that minimize expected free energy), this seems to imply that intentions are reducible to belief-desire pairs. As argued above, this clashes with accounts that attribute distinctive and irreducible roles to intentions. Others maintain that the belief-desire account can explain all aspects of intentions (Sinhababu 2013). We do not intend to settle this complex issue. Our aim is simply to clarify what theories of action are available within a preference PP framework.

Revisionist aspirations are considered important in some PP circles. However, such aspirations are optional within a preference PP framework. Preference PP does not require any major revisions to standard conceptions of motivation and action. In this respect, preference PP deviates little from other prominent computational frameworks, such as reinforcement learning and Bayesian decision theory, which also contain distinct representations and computations of value. Thus, preference PP is potentially much less revisionist than its predecessor, optimistic PP, and does not entail the radically revisionist program advocated by some theorists. Preference PP might aspire to offer a universal account of agency, and, perhaps, the mind in general, but this universality is traded off against the supposed simplicity of the framework and its formalisms.

6. What about the Free Energy Principle?

6.1. The low road and the high road

So far, we have been pursuing what some PP theorists have called the low road to PP (Friston, 2019; Parr et al., 2022, Chap. 2). The low road starts from the assumption that the brain is a Bayesian inference engine trying to optimize its model of the causes of its sensory input. PP is then proposed as an explanation of how the brain can solve this inferential problem and the neurocognitive mechanisms involved in this process. By enriching PP models with whatever constructs necessary to explain the empirical data, PP might gradually come to explain more and more aspects of cognition and behavior, including action, motivation, planning, and decision-making.

The low road is sometimes contrasted with the high road to PP. The high road takes as its starting point fundamental questions about what properties systems that manage to persist must have. According to Friston (2019, p. 177), 'any system that exists will appear to model and predict its exchange with the environment'. More specifically, any self-organizing system will necessarily engage in (or necessarily appear to engage in) the minimization of free energy. This idea is known as the *free energy principle* (FEP). According to proponents of the high road, PP turns out to be a necessary feature of self-organizing systems. For this reason, the high road has been described as a 'top-down journey from near existential nihilism to the riches of predictive processing' (Friston, 2019, p. 175).

Let us unpack the FEP. A living organism must resist a tendency to disintegrate: it must keep its internal states within a viable range as reflected by their homeostatic properties. To do so, it must conserve a boundary that distinguishes it from its environment. Under the FEP, this boundary is formalized as a Markov blanket (Friston, 2013). Markov blankets are supposed to partition states into those internal to the system, external to the system, and the states of the boundary itself. Some boundary states are influenced by external states (namely, sensory states) and some by internal states (namely, active states). States that, according to the organism's model of the world, are expected to be incompatible with its continued existence are deemed surprising (in a technical sense). These states are deemed unlikely to occur when the organism inhabits a hospitable environment. Since calculating surprise directly would require knowing all the hidden states of the world that cause the sensory input, it is impossible for any organism to calculate this directly. Instead, it must minimize variational free energy, which is an upper bound on surprise. The organism thus effectively minimizes surprise in the only tractable manner. In other words, living systems expect to occupy states compatible with their continued existence. By minimizing variational free energy, the system keeps its internal states within a range consistent with its survival.[12]

In short, the high road involves developing so-called process theories that align with the FEP. These theories account for the structure and functions of neurocognitive mechanisms, which are consistent with the imperative of minimizing free energy. Under this approach, PP is essentially the suite of such process theories. It is important to emphasize that preference PP models, which explain policy selection in terms of expected free energy minimization, are not committed to the broader claims of the FEP. Variational free energy and the expected free

[12] As some authors have argued (Seth, 2015; Pezzulo & Cisek 2016), this makes the FEP a modern version of cybernetics, according to which control consists in using feedback signals to keep essential internal variables within an expected range.

energy of policies are not the same. Variational free energy minimization serves to model and predict the environment based on past and present observations. Expected free energy minimization, by contrast, pertains to action selection based on expectations about the consequences of future actions (see Parr et al., 2022, pp. 31–39).

6.2.The desert landscape

Why discuss the FEP? Because some claim that the FEP entails a desert landscape view of the mind: a minimalist ontology in which 'there are neither goals nor reward signals as such' (Clark, 2013, p. 200). All that really exists is self-organizing systems that appear to model and predict their environment via free energy minimization (Friston, 2019). Although resisted by Clark (2013), this view is (at least sometimes) endorsed by Friston (2019) and other proponents of the FEP (Ramstead et al., 2019) and presented as an inevitable consequence of the FEP. The desert landscape interpretation of the FEP denies the existence of pro-attitudinal constructs, such as value, reward, goals, drives, and desires. Therefore, if this radical interpretation is true, then PP (in any guise) is necessarily anti-Humean under the FEP.

There has been much disagreement about the high road that takes the FEP as its theoretical starting point, and what it entails exactly (Clark, 2013; Friston, 2019; Williams, 2022). For our purposes, the relevant question is whether the FEP entails a desert landscape ontology, which would restrict process theories to anti-Humean varieties with no pro-attitudinal constructs. As we will see, this depends on how the FEP is interpreted.

On one reading, the FEP strives to explain how self-organizing systems manage to persist over time by means of mechanisms that implement free energy minimization. If free energy minimization is a necessary condition on self-organizing systems, and any mechanism implementing it precludes pro-attitudinal states, then this would entail a desert landscape view of the mind.

On another reading, the FEP merely posits that all self-organizing system can be redescribed as if they minimize free energy. Under such a reading, the FEP places no constraints on how such systems minimize free energy, and process theories are free to include pro-attitudinal constructs. The specific mechanisms involved could be very different for rocks, oil drops, and humans. Under this interpretation, PP process theories could include pro-attitudinal constructs, so long as it remains true that the target system can be described as it if it minimizes free energy—even if this is not strictly the objective of the mechanisms underlying

the system's behavior. This interpretation does not entail a desert landscape ontology and is consistent with Humeanism.[13]

Others have argued that the FEP is consistent with a folk-psychological distinction between beliefs and desires. Smith et al. (2022) argue that there are terms within the expected free energy formalism (that is, within preference PP), which can be functionally identified with desire-like constructs with a world-to-mind direction of fit as described by folk psychology.[14] Even when described by the FEP, they argue, the organism can still be described as having desires. This illustrates that there is no clear consensus that the FEP entails a desert landscape ontology. For those who deny this implication of the FEP, there need be no conflict between the FEP and a Humean interpretation of preference PP.

7. Conclusion

We have explored the intricacies of the predictive processing framework by uncovering two distinct theories within it and their distinct implications for motivation and action. The difference between these has not been properly appreciated. Optimistic PP bases all processing on optimistic priors and entails a revision of standard accounts of motivation and action. This gives rise to significant explanatory challenges. Sticking to its simplistic formalism, optimistic PP must relinquish its ambition to provide a universal account of the mind. There are aspects of motivation and action that seem beyond its explanatory scope. By contrast, preference PP posits that actions are selected to minimize expected free energy and aligns more closely with standard accounts of motivation and action in philosophy and cognitive science.

Contrary to some radical interpretations, the FEP does not necessitate a fundamental overhaul of standard desire-like or pro-attitudinal constructs. Preference PP explains more aspects of motivation and action. However, the broader explanatory scope requires moving beyond attempts to reduce all mental phenomena to a single process of Bayesian belief updating. In its preference PP incarnation, PP has instead come to resemble other prominent computational frameworks implementing a distinction between beliefs and desires, such as reinforcement learning and Bayesian decision theory. The general lesson is that a tension exists

[13] For an argument that the FEP places no necessary constraints on explanations of how self-organizing systems manage to maintain their existence, see (Williams, 2022).

[14] Though similar in some respects, our analysis is different in others. One difference is that we focus on the role of desire-like constructs in philosophical and scientific theories of motivation and action, not simply on consistency with how they are described in folk psychology. Another is our focus on value learning and the need for a non-Bayesian account thereof.

between the parsimony often aspired to in PP theories and accepting enough primitives to give a complete account of agency and the mind.

References

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function, 218*:611–643.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018a). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese, 195*:2417–2444.

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018b). Free-energy minimization in joint agent–environment systems: A niche construction perspective. *Journal of Theoretical Biology, 455*:161–178.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*:181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and Cognition, 56*:115–119.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy, 98*:1–15.

Colombo, M. (2017). Social motivation in computational neuroscience. In J. Kiverstein (ed.), *The Routledge Handbook of Philosophy of the Social Mind* (pp. 320–340). Routledge.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy, 60*:685–700.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*:293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*:127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface, 10*:20130475.

Friston, K. J. (2013). Active inference and free energy. *Behavioral and Brain Sciences, 36*:212–213.

Friston, K. J. (2019). Beyond the desert landscape. In M. Colombo, E. Irvine, & M. Stapleton (eds.), *Andy Clark and His Critics* (pp. 174–190). Oxford University Press.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS One, 4*:e6421.

Friston, K. J., Samothrakis, S., & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics, 106*:523–541.

Friston, K. J., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark room problem. Frontiers in Psychology, 3:130.

Friston, K. J., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience, 7*:598.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6*:187–214.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation, 29*:1–49.

Friston, K. J., Salvatori, T., Isomura, T., Tschantz, A., Kiefer, A., Verbelen, T., Koudahl, M., Paul, A., Parr, T., Razi, A., Kagan, B. J., Buckley, C. L., & Ramstead, M. J. D. (2025). Active Inference and Intentional Behaviour. *Neural Computation*, *37*(4):666–700.

Grünbaum, T., & Christensen, M. S. (2024). The functional role of conscious sensation of movement. *Neuroscience and Biobehavioral Reviews, 164*:105813.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*:259–285.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press.

Klein, C. (2018). What do predictive coders want? *Synthese, 195*:2541–2557.

Klein, C. (2020). A Humean challenge to predictive coding. In D. Mendonça, M. Curado, & S. S. Gouveia (eds.), *The Philosophy and Science of Predictive Processing* (pp. 25–38). Bloomsbury.

Lewis, D. K. (1988). Desire as belief. *Mind, 97*:323–332.

Lewis, D. K. (1996). Desire as belief II. *Mind, 105*:303–313.

Miller Tate, A. J. (2021). A predictive processing theory of motivation. *Synthese, 198*:4493–4521.

Murphy, K. P. (2023). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports, 7*:14678.

Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology, 29*:1 6.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20*:414–424.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation, and adaptive behavioural control. *Progress in Neurobiology, 134*:17–35.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences, 22*:294–306.

Ramstead, M. J., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews, 31*:188–205.

Ransom, M., Fazelpour, S., & Mole, C. (2017). Attention in the predictive mind. *Consciousness and Cognition, 47*:99–112.

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation, 33*:674–712.

Schueler, G. F. (1995). *Desire: Its Role in Practical Reason and the Explanation of Action*. MIT Press.

Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing*. MIND Group.

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences, 24*:681–683.

Sinhababu, N. (2013). The desire–belief account of intention explains everything. *Noûs, 47*:680–696.

Smith, M. (1987). The Humean Theory of Motivation. *Mind, 96*:36–61.

Smith, R., Ramstead, M. J. D., & Kiefer, A. (2022). Active inference models do not contradict folk psychology. *Synthese, 200*:79.

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences, 24*:346–348.

Van de Cruys, S., Friston, K. J., & Clark, A. (2020). Controlled optimism: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences, 24*:680–681.

Wiese, W. (2017). Action is enabled by systematic misrepresentations. *Erkenntnis, 82*:1233

1252.

Williams, B. (1979). Internal and external reasons. In R. Harrison (ed.), *Rational Action* (pp. 101–113). Cambridge University Press.

Williams, D. (2022). Is the brain an organ for free energy minimisation? *Philosophical Studies, 179*:1693–1714.

Yon, D., de Lange, F. P., & Press, C. (2019). The predictive brain as a stubborn scientist. *Trends in Cognitive Sciences, 23*:6–8.

Yon, D., Heyes, C., & Press, C. (2020). Beliefs and desires in the predictive brain. *Nature Communications, 11*:4250.

Concluding Remarks

The aim of this dissertation has been to advance our understanding of thought. Rather than being passive and unguided, mind-wandering is an actively guided process that supports planning and learning. Cognitive maps mediate mental transitions that plausibly qualify as reasoning, without taking the form of rule-governed operations over propositional attitudes. A closer examination of the predictive processing framework underscores that the distinction between beliefs and desires—standard in much of philosophy and cognitive science—remains indispensable for a complete account of agency and mind. In this way, the dissertation both challenges traditional boundaries of rational and active thought, broadening our understanding of the diverse ways we conduct our mental lives and affirms certain established views about mental function while suggesting ways to refine them.

Although each article can be read as a self-contained project, taken together they yield broader lessons. The conclusions of the articles are complementary in several respects. While Article 1 highlights how mind-wandering supports planning, Article 2 further examines the guidance mechanisms that enable it to do so. Article 2 shows that mind-wandering facilitates learning, and Article 3 shows how transitions of the sort underlying such learning processes can qualify as reasoning. In turn, Article 2 demonstrates how transitions like those underpinning map-mediated reasoning can be actively guided. Finally, Article 4 highlights that distinct belief-like and desire-like states are needed to explain action and motivation—a point that plausibly extends to mental action.

While the dissertation has focused on what we can learn from studying specific mental phenomena, the results also bear on broader questions about the nature of the mind. The dissertation generally supports the representational theory of the mind: it shows how a variety of representational and computational mechanisms are needed to explain our capacities to think, reason, and act. According to representationalism, mental processes are causal processes involving the interaction of physical entities—representations—that carry content (Fodor, 1987; Dretske, 1988; Millikan, 1989; Burge, 2010; Shea, 2018). Mental representations are internal, content-bearing structures that refer to things in the world. Mental states such as beliefs, desires, intentions, thoughts, perceptions, or imaginings are treated as relations to mental representations. For example, to believe *P* is to stand in an appropriate relation to a mental representation with *P* as its content (Quilty-Dunn & Mandelbaum, 2018). Similarly, to desire *P* is to stand in a different relation to the same mental representation. Mental processes such as thinking, reasoning, or imagining are explained as causal sequences of mental

representations that are, in some way, realized in the brain. Representationalism continues to be the dominant theory of the mind within philosophy, and its core assumptions are implicit in most areas of cognitive science.

A prominent alternative to representationalism is dispositionalism, which holds that beliefs and desires are dispositions to act, feel, and think in certain ways.[1] On this view, beliefs and desires are defined by distinct dispositional profiles. To believe *P* is to be disposed to act, reason, and react in ways consistent with *P* being true. For example, believing it is raining disposes you to carry an umbrella, comment on the weather, or seek shelter. To desire *P* is to be disposed to pursue actions or exhibit motivational responses aimed at bringing about *P*. For example, desiring a drink disposes you to reach for a glass, seek water, or feel satisfaction if you drink. Dispositionalists reject the representationalist view that beliefs and desires are relations to internal representations carrying their content, and that the effects of beliefs and desires are explained by causal processes involving those representations.

While some contend that predictive processing supports internal, content-bearing representations (Hohwy, 2013; Clark, 2016), others resist this view and opt for non-representational interpretations of the framework (Bruineberg et al., 2018; Ramstead et al., 2019). So, while predictive processing is compatible with representationalism, the former does not seem to entail the latter. The formalism of predictive processing is not inherently committed to the existence of internal, content-bearing representations but can instead be interpreted as a formal description of dispositional patterns. Within the formalism of expected free energy models (preference predictive processing), the dispositional profile of beliefs and desires might be described by distinct formal terms: beliefs are dispositions to reduce uncertainty, and desires are dispositions to pursue preferred outcomes.

However, dispositionalism sits poorly with the other results of the dissertation. Mind-wandering and map-mediated reasoning seem to involve causal processes over content-bearing entities in the mind and brain. What is being monitored, evaluated, and regulated during rational and active thought are a variety of representational structures and the causal operations performed over these structures. These structures include sensory, motoric, affective, evaluative, and map-like representations, each activated by special-purpose systems. When these representations are deemed relevant to an ongoing mental process, they attract attention and can be accessed by working memory, where they are integrated into complex mental

[1] For a defense of dispositionalism about beliefs and a critique of representationalism, see Schwitzgebel (2002, forthcoming). For a dispositionalist theory of desires, see Smith (1987).

simulations. These simulations allow us to evaluate different scenarios in terms of their likelihood, value, and costs. On this basis, we draw conclusions about what is the case and how to act. The discovery of all these mechanisms and their interplay pushes toward a representationalist reading of the frameworks we use to model mental phenomena. For example, on a representationalist reading of predictive processing, preferred outcomes could be understood as representations of the value of actions or outcomes. This interpretation aligns with recent accounts inspired by other computational frameworks, such as reinforcement learning and neuroeconomics, which treat desire-like states as value representations (Railton, 2017; Haas, 2023; Carruthers, 2025; Sripada, 2025).

A full appreciation of the diverse representational and computational mechanisms that constitute the mind also compels us to move beyond purely logic-based notions of rationality. While conclusions are sometimes reached by following broadly-logical rules, non-logical transitions between thoughts are ubiquitous and often result in reasonable conclusions. These non-logical, content-specific transitions are often responsive to rational norms, tracking the quality and costs of different strategies, recruiting the representational structures best suited to the task at hand, and selecting responses expected to have the best outcomes. Cognitive maps play an important role in mediating such content-specific transitions: they structure the development of mental simulations and guide attention toward relevant content. Given their ability to generate conclusions in ways that are responsive to rational norms, these content-specific transitions should be recognized as rational forms of reasoning.

In conclusion, this dissertation advances a pluralistic view of thought—one that recognizes the many ways we conduct our mental lives in the pursuit of our goals. It contributes to ongoing efforts to map the architecture of the mind by offering insights into how we plan, reason, and exercise control over our thinking. An important avenue for future research lies in further clarifying how interactions among diverse representational structures guide thinking, reasoning, and action. This project is already underway, yet the complexity of the mind and brain ensures that much remains to be done. That complexity should not be simplified away: the mind is an intricate machine with a diversity of tools at its disposal that are themselves quite complex. Confronting this complexity will almost certainly require revising or abandoning old hypotheses when they prove inadequate for capturing the riches of the mind. However modestly, I hope this dissertation has taken a few steps toward that broader goal.

References

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6):2417–2444.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press.

Carruthers, P. (2025). *Explaining our Actions: A Critique of Common-Sense Theorizing*. Cambridge University Press.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Dretske, F. I. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Haas, J. (2023). The evaluative mind. In J. Haugeland, C. F. Craver, & C. Klein (eds.), *Mind Design III: Philosophy, Psychology, and Artificial Intelligence* (pp. 295-313). MIT Press

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Millikan, R. (1989). Biosemantics. *The Journal of Philosophy, 86*(6):281-297.

Quilty-Dunn, J. & Mandelbaum, E. (2018). Against dispositionalism: belief in cognitive science. *Philosophical Studies, 175*(9):2353-2372.

Railton, P. (2017). At the Core of Our Capacity to Act for a Reason: The Affective System and Evaluative Model-Based Learning and Control. *Emotion Review*, *9*(4):335-342.

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, *28*(4):225–239.

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. *Noûs, 36*(2):249-275.

Schwitzgebel, E. (forthcoming). Dispositionalism, yay! Representationalism, boo! In J. Jong, & E. Schwitzgebel (eds.), *The Nature of Belief*. Oxford University Press.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Smith, M. (1987). The Humean Theory of Motivation. *Mind, 96*:36–61.

Sripada, C. (2025). The valuationist model of human agent architecture. *Philosophical Psychology*, 1–30.