Rational Structures in Learning and Memory

by

Sara Aronowitz

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2018

Doctoral Committee:

Professor Peter Railton, Chair
Professor James Joyce
Professor Richard Lewis
Associate Professor Chandra Sripada
Professor Timothy Williamson, Oxford University

A frame from Yuri Norshteyn's film *The Tale of Tales*. 1979

Sara Aronowitz

skaron@umich.edu

ORCID iD: 0000-0001-9099-3031

ACKNOWLEDGEMENTS

I can't convey my immense gratitude here towards everyone who has helped and guided me during the writing of this dissertation, so I'll instead be brief. I was lucky to start grad school along with Mara Bollard, Sydney Keough and Sophie Monahan. Among many other things, I am grateful to Sophie for exploring the most profound and silliest imaginative journeys, to Mara for sharing her ideas which always cut so precisely to what is most important, and to Sydney for more than I could begin to put into words. I have benefited from working with basically everyone who I overlapped with at Michigan. In particular, Daniel Drucker and Eduardo Martinez helped me with drafts, Nina Windgaetter was an ideal work companion, and Alice Kelley and Robin Zheng collaborated with me.

My advisor, Peter Railton, taught me so much, and was unfailingly patient, kind, and willing to entertain and discuss even my most insane ideas. I never felt like grad school was an exercise in getting a job or impressing people, and that was because of Peter. My incredible committee members Jim Joyce, Chandra Sripada, Rick Lewis, and Tim Williamson helped shape (the good parts of) this dissertation down to the small details, and introduced me to countless new ideas as well as critical features of old ideas that I had missed. Ishani Maitra and Brian Weatherson gave me invaluable feedback on drafts. I got to take so many wonderful courses at Michigan, including those taught by Gordon Belot, Sarah Buss, Victor Caston, Dan Jacobson, Maria Lasonen-Aarnio, Sarah Moss, Gina Poe, Thad Polk, Jamie Tappendan and Rich Thomason. Victor Kumar was the best possible boss. This dissertation somehow ended up including a short discussion of classical Islamic philosophy, which I would never have learned about had I not been lucky enough to study with Deborah

Black.

As a visiting student at Rutgers, I got so much invaluable advice and feedback from Susanna Schellenberg - and I was lucky enough to get to see the inner workings of some of her fascinating philosophical theories. Randy Gallistel made time for me pick his brain about all kinds of topics, and Pernille Hemmer welcomed me into her lab. Ernie Sosa let me sit in on his amazing seminar. Branden Fitelson always got me excited about philosophy. Liz Camp is the kind of scholar that I aim to be someday. So many Rutgers grad students invited me in to their work- and non-work-related activities.

I have presented parts of this dissertation at many conferences and workshops and received feedback from several anonymous referees. I'd also like to thank my conference-buddies, who I got to know in sometimes bizarre circumstances, in particular Marilie Coetsee and Amir Saemi. A walk with Sasha Nabokov is irreplaceable. And of course, thanks to Reza Hadisi for everything.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

ABSTRACT vii

CHAPTER

I. Introduction 1

II. Memory is a Modeling System 3

2.1 From Storage to Retrieval 6
2.2 Sleep and Memory 10
2.2.1 Predictions 11
2.2.2 Data 12
2.3 Two Retrieval Systems 16
2.3.1 The Model System 17
2.3.2 The Index System 21
2.4 Comparing the Index and Model Systems 23
2.5 Epistemological Context 26
2.5.1 Causal Theories 26
2.5.2 Memory-as-Testimony 30
2.6 Wrong Level of Abstraction? 31
2.7 Conclusion 32

III. Exploring by Believing 35

3.1 The Exploration/Exploitation Trade-Off 37
3.1.1 Prelude 38
3.1.2 The Multi-Armed Bandit 39
3.1.3 Extending the Trade-Off 43
3.2 A Case of Exploratory Belief 48

3.2.1 Why Belief? . . . 48
3.2.2 The Example . . . 48
3.3 Argument for a Belief/Action Symmetry . . . 52
3.4 Conditions 1 and 2 . . . 55
3.5 Condition 3 . . . 56
3.5.1 Belief and Evidence-Gathering . . . 56
3.5.2 Consideration and New Hypotheses . . . 59
3.6 Condition 4 . . . 62
3.7 Objections, Alternatives . . . 63
3.8 Conclusion . . . 67

IV. Memory Anomalies . . . 69

4.1 Cognition of Anomalies . . . 73
4.1.1 The Big Picture . . . 74
4.1.2 Novelty in Memory . . . 75
4.2 A Distinctive Function . . . 79
4.3 A Rational Function . . . 81
4.3.1 Closer to Truth . . . 82
4.3.2 Apt Response to Evidence . . . 84
4.3.3 Coherence . . . 85
4.4 Upshots, and a Metaphor . . . 89

V. Belief, and its Fraying Edges . . . 92

5.1 Introduction . . . 92
5.2 Belief-like Memory . . . 93
5.3 The Problem, and Some Partial Solutions . . . 97
5.4 The Al-Farabian View . . . 102
5.4.1 Conceptualization and Generality . . . 107
5.5 Returning to the Cases . . . 110
5.5.1 Comparing the Taxonomies . . . 113
5.6 You Only Are What You Believe? . . . 115

VI. Conclusion . . . 120

BIBLIOGRAPHY . . . 123

LIST OF FIGURES

Figure

3.1 A simple bandit problem, where arms have a bias drawn from an underlying distribution and fixed noise 41

3.2 An epistemological version of the bandit problem, where choosing an arm represents changing your belief state. Just as the agent in Figure 3.1 rationally alternates between the arms even though endlessly repeating the highest estimated arm has the maximum myopic value, Nurudin rationally samples from all of the arms in this figure, or perhaps alternates between j and k, rather than picking the myopically optimal arm j over and over. 52

3.3 Nurudin's position on the proposition in question will lead him to consider and form beliefs about different sets of propositions, whose truth is related in complex ways to the truth of the original proposition. 55

3.4 Nurudin's wandering consideration. As he explores the space of possibilities from two different starting points (rectangles), he encounters different propositions of the sort listed in Fig. 3, depicted here as stars. The process of moving around this space has some degree of randomness, but also reflects the generalization that positions nearer to starting points are easier to explore and attract more attention, ceteris paribus. 61

3.5 Sketch of possible relations between beliefs and imaginative search 61

4.1 An exhibit from the Museum of the Great Patriotic War in Moscow, depicting the siege of Leningrad. 89

4.2 An online exhibit highlights the puzzles surrounding the fall of the city of Teotihuacan, and draws attention to a reconstructed artifact (right) which was apparently broken up into many pieces and deliberately separated across many disparate locations 90

ABSTRACT

My dissertation aims to disrupt an increasingly ubiquitous view of epistemology which claim that we can study rationality by considering a single belief at a single time. I target three areas where diachronic (i.e. temporal) factors make a difference in the three sections:

1. memory, a system of tremendous importance in our cognitive lives yet which is often reduced to a one-sided question of whether to trust what one's memory says,

2. learning, where I argue that we should sometimes believe in a way that's not warranted or reasonable in light of our current evidence, but which puts us in a better position to acquire evidence in the future, and

3. the connection between memory and learning, as exemplified in the case of remembering anomalous events.

While many disciplines dealing with memory have come to the view that it is not a passive system for recording and preserving the data of experience, most extant philosophical accounts implicitly or explicitly make the assumption that storage is nonetheless memory's core function - other characteristics are secondary or derivative. In Chapter II, I aim to restructure the problem, moving the focus from storage to retrieval the ability to pull up relevant and accurate information on demand. On my view, the memory system is different from active mental processes such as deliberation only by degree (i.e., it is slower to update) and in virtue of how it is connected with other faculties (memory tends to be updated

off-line, rather than in concert with decision-making and perception). I call this the model' view of memory.

In Chapter III, I extend a decision-theoretic concept (the exploration-exploitation trade-off) to the case of belief. The trade-off is between picking the best-looking option, and picking the option most likely to lead to learning. For belief, this connection is grounded in the connection between our current beliefs and our dispositions to act in more or less experimental ways and to explore the space of possibilities (in a process that I call imaginative search). These acts - experimentation and imaginative search - change what evidence we have, and yet at the same time also depend causally and rationally on what we believe. We perform one experiment over another because of our beliefs about what is probably true of the system we're studying, and what the result of the experiment would tell us about that system. The same goes for imagination, albeit more controversially.

In Chapters IV and V, I analyze the rationality of selective memory for unexpected events. Chapter IV surveys empirical work on anomalous memories, in order to understand the rationality of the cognitive process underlying these curious states . Where does remembering anomalous events fit into the behavior of an epistemically responsible agent? I argue that generating and maintaining anomalous memories is a distinctive rational function of the long term memory system. This claim has two parts. First, this kind of cognitive behavior is epistemically rational. Second, it is a cognitive operation that is particular to the memory system, meaning that the operation itself is a memory operation as opposed to a computation that just happens to take place in the memory system. And so this conclusion entails that memory itself has a distinctive, active epistemic role. Chapter V uses the case of anomalous memories as a jumping-off point to ask: do we believe everything we remember? I resurrect a medieval view of mental states, based on a theory advanced by Abu Nasr Al-Farabi, which factors our belief-like states into two components: conceptualization and commitment. On this analysis, the thing we typically call belief is a small corner of a larger space that includes the variety of ways we build memory representations

and use those representations to guide us.

This project is important because our whole lives are organized around getting things right at the right time. When we try to act morally, we might try to have a life that is built around moral principles, or to become wiser and kinder over time, as opposed to amassing a collection of acts that all have independent moral value. I think the same thing is true of our endeavors to acquire knowledge the process of inquiry is not made up of individual, independent good inferences that happen to follow one another, but is instead about a trajectory where we learn over time, and take the right steps now to get things right in the future, and overall. So I think that to understand this more complete sense of inquiry, philosophy needs to make a place for memory, the system that sustains and directs inquiry in the background, over long periods of time even as the sciences are learning more and more about how natural memory systems work, philosophers have boxed it out of relevance.My methodology is to study natural and artificial learning and memory systems as a process of discovery, a way of using real-world cases as inspiration and guide to the normative landscape. Conversely, I hope that figuring out new normative possibilities can shed light on empirical facts - though this is not the main focus of my dissertation.

CHAPTER I

Introduction

I started the project that would become this dissertation by thinking about reversal learning in rodents. In a simple learning task that involves this skill, rats have learned the location of an underwater platform, and are now quite successful at swimming to that platform when put into the tank. Then, the location of the platform is changed, and the rat must find the new location on subsequent trials. This ability, to identify when old information is no longer relevant and replace it with new information, is linked to REM sleep; rats that are REM-deprived do much worse at adapting to the new location, and are more likely to return to the old location. Studying the neural processing behind this ability to re-learn, something began to bother me.

There is something more complex behind reversal learning. The rat does not just learn that the platform was here and now it is there. It also reconfigures its behavior in response to this change, so that it will swim in the right direction. This in turn involves re-working an existing model of the environment, and the rat's own body, to plot a new route. What began to bother me is that it seems like this kind of learning requires sleep for a reason - roughly, because undoing existing connections and moving things around is hard to do when you are actively relying on the model for behavior. But my philosophical toolbox did not offer any way to explain what is special about this kind of learning, or why it naturally takes place off-line. Most of the epistemology I was familiar with was about

learning the generic proposition P, and didn't seem to be sensitive to differences between the types of things being learned. Likewise, most philosophical models of learning that I was familiar with were insensitive to *when* learning should take place, and ignored the possibility of learning during sleep. This bothered me because it seemed like there was actually something normatively significant about reversal learning, about the differences between learning and un-learning, and about the trajectory of learning the environment across time and changes in routine and cognition.

You might be unconvinced that any of this is actually of philosophical significance. But the aim of this dissertation is to change your mind, and to provide a theoretical, normative foundation for thinking about features of learning, un-learning and re-learning over long timespans. These chapters are meant to stand on their own, but together they build an epistemology that is attentive to the thinker as an agent moving through time and facing a series of related challenges. The picture that emerges of epistemic rationality deviates from the orthodoxy in various ways - which is to say that the decision to focus on a single point in time or an extended trajectory is *normatively significant*.

CHAPTER II

Memory is a Modeling System

ABSTRACT. This paper aims to reconfigure the place of memory in epistemology. I start by rethinking the problem that memory systems solve; rather than merely functioning to store information, I argue that the core function of any memory system is to support accurate and relevant retrieval. This problem formulation has consequences for what structures and mechanisms make up a memory system. In brief, memory systems are modeling systems. This means that they generate, update and manage a series of overlapping, simplified, relational representations that map out features of the world. Succeeding at building and maintaining models requires the kind of active knowledge generation traditionally associated only with deliberative reasoning.

Introduction

I'll start with an analogy. Some philosophers think that memory works like a museum. It's a place where old things are collected and stored over time. These old things are important not just because they are interesting and informative, but also because they are authentic links to our shared past. The function of the museum is simple: be a receptacle for these old things. A good museum is the right temperature so that the artifacts don't degrade - and likewise, on this picture, a good memory system maintains a cognitive environment to keep the items stored in memory in as close to their original condition as possible. In any real museum, items do degrade and need to be touched up and patched. But the idea of keeping things in the original condition is a guiding norm. Similarly, while human memory might involve alterations to the stored information, the system is still aiming at or guided

by the norm of perfect preservation. A recent trend, constructivism, rejects the museum picture entirely. For instance, Michaelian [65] makes the case that memory just is a kind of imagining - it has no special relationship to preserved content at all.

On the view I develop in this paper, memory is indeed analogous to a museum, though not in the way I've just described. To explain, I'll sketch a different picture of how museums function. While they do usually contain some old and authentic artifacts, there is more to a good museum than good preservation. For one, there's a team of curators who arrange exhibits. This involves choosing what to display - but perhaps more importantly, organizing a series of objects around a theme or narrative. For instance, a curator might plan an exhibit about Turgenev. She would choose an angle - say, exploring his literary circle and emphasizing connections between France, Russia and England. The choice of a theme will depend on the artifacts at her disposal, but it also will determine which she chooses to display. The same goes for the process of arranging the exhibit and generating displays like maps, captions and installations. In the process of making a map, she'll realize it would be helpful to display a postcard from Turgenev's travel to Italy or in the process of arranging a set of letters, she'll realize it would be helpful to have a chart of his correspondence. A museum, in this sense, is an active institution as opposed to a passive repository. And this is most obvious when we consider the distinctive job of curation that occurs in museums.

Thus on this analogy, we can position alternative views of the computational function of memory on a scale from memory having the most active role in information operations to the most passive role. On the first extreme, the exhibits in the museum might be purely generated by the curators, incorporating no received artifacts and relying on minimal external resources. This view is of course *memory constructivism*. On the other extreme, the traditional *memory preservationism*, the curators merely carry out instructions given to them from outside.

In this paper, I'll argue that memory is somewhere between these two extremes, and operates a lot like an actual museum. In particular, the memory system performs an epistemic

activity which is analogous to curation, which I'll call modeling. This operation is aimed at structuring and altering its contents in order to make correct, useful and relevant information available for retrieval. I'll contrast this picture to both the initial museum analogy, where memory is aimed at authenticity and preservation, and to the constructivist account on which memory is a system of imaginative simulation. Just as curatorial work is mainly done when the museum is closed to visitors, the memory operations I'm interested in occur during sleep and in the background during waking, rather than being conscious, online processes.

This paper aims to answer two questions in the epistemology of memory: (1) what is the epistemic role of memory, i.e. the role of memory in generating knowledge? and (2) which features of the memory system are epistemically evaluable? My answer to these normative questions is inspired by recent work in computer science and neuroscience that has allowed us to dissect natural and artificial memory systems and better understand the problems which they succeed at solving.

While this account is about memory in general, I make several assumptions about the evidential context of the agent in question: (a) the agent has a set of evidence which is small relative to the total possible evidence about propositions which interest the agent, (b) the agent can expect to get more information over time that (at least in part) bears on the same propositions, and (c) the agent is handling a large amount of complex information. Roughly, human memory falls into a class of systems solving a common problem: how to use large quantities of information under conditions of uncertainty when you expect to accumulate more and more in the future. This could apply to humans or other agents such as rats or primates, as well as machines that fall short of full agency, such as a database or specialized artificial intelligence.

Conditions (a)-(c) are plausible claims about most of our cognitive circumstances, but they are meant to capture something more specific: namely, what it means to be in the beginning of inquiry. I'll argue that given these conditions, memory isn't just a system for

storing information, but actually has its own procedures and norms for generating new content and revising old content. On this picture, memory is differentiated from deliberation only by its degree of holistic connection and overall conservativeness.

The structure of this paper is as follows: in Section 2.1, I lay out the claim that retrieval is central to memory, and support it with empirical evidence that suggests retrieval is the most significant bottleneck in successful memory function. The museum analogy suggests that our memory systems are doing work in the background to update and change stored information; in Section 2.2, I show that active information transformation is happening during sleep in humans and other animals based on neuroscientific research. Understanding which transformations are happening and why requires a conceptual model. In Section 2.3, I describe two informational structures that solve the retrieval problem - first, a model system that structures the data in a series of simplified models, and second, an index system, which uses an intermediate structured representation between the query and the underlying unstructured storage system. I define these two types and present instances of real-world model and index devices. In Section 2.4, I argue that the model system is a better fit for memory in an ideal agent, by showing that as computational constraints are relaxed, index systems transform into model systems. In Section 2.5, I survey some views of memory in the philosophical literature, and show that they are problematic given the norm described in Section 2.1. I conclude by considering an objection that the model view isn't at the right level of abstraction, and discussing some consequences of the model view.

2.1 From Storage to Retrieval

Here's a seemingly trivial fact: memory systems must make their stored information available for retrieval. However, this fact that has been given short shrift in the current epistemological literature. If we think of memory as a process that begins with some experience and leads to a later experience of remembering, most philosophers have focused on the experience of remembering itself. Memory constructivists like Michaelian[65], for

instance, are interested in how we are able to fill out the details and construct a model of the past when having the experience of remembering, or shortly before. On the other hand, making stored information available for retrieval is a process that starts as soon as we have an experience, and is ongoing until the experience of remembering and afterwards. This is a process over days, months or years which determines what gets remembered; the constructivist is concerned with a process that takes minutes which determines how the experience of remembering is generated and experienced.

What is retrieval? It's a process in which a system supplies an item or set of items that relate to a query. Queries can be explicit - such as when I deliberately search my memories to find my brother's girlfriend's name - or implicit - such as when seeing a person's face cues a memory of their name. Retrieval in fact involves two processes working together. In the long run, stored information must be structured so that important things can be retrieved. In the short run, search has to operate on stored representations successfully to locate relevant stored content.

Retrieval is a functional bottleneck for memory, a step in the process to which many failures and successes can be traced. This holds for both animals and artificial systems. In using this fact to argue for the normative centrality of retrieval, I'll assume that one of the theoretical virtues that makes for a good account of the function of a biological system is that it can explain common successes and failures.

Research on human memory suggests that retrieval is a major source of memory failures[88]. One way of seeing this intuitively is to think of all the times you couldn't remember some fact that later came to you easily in a different context.In computer memory systems, where it's fairly easy to tack on another server or hard disk or some other storage device, the retrieval problem is even more prominent; consider how much improvements in web search facilitate our use of information. One could argue this is just a feature of the peculiar strengths and weaknesses in our technology, but it's significant insofar as it aligns with the challenges found in biological memory systems. The fact that retrieval shows up as

the factor which separates successful from unsuccessful memory systems across the board suggests that retrieval is a challenging computation even at a high level of abstraction.

Strategies to enhance retrieval are central to memory expertise in humans. The Method of Loci (MoL) is a strategy used by memory experts as far back as Roman times[25], and notably described by A.R. Luria as used by the brilliant mnemonist Shereshevsky[58]. An example of this technique would be memorizing a list of unrelated names by visualizing placing each name in a familiar location on a walk around my childhood neighborhood. Then, to retrieve the items, I would again visualize my walk through the neighborhood, this time picking up each name one by one. The MoL is not just used by experts but can be taught to older adults to improve memory performance[45][30]. The MoL is a retrieval strategy because it involves structuring one's memories so that they can be effectively retrieved later, and prescribes a specific strategy for that retrieval process. The symmetry between how the memory is stored (in a visualization of the childhood home) and then consequently retrieved is striking, and illustrates the property of retrieval I discussed above: effective retrieval involves tight cooperation between how information is stored, and how it is searched.

Another line of evidence is retrieval pathologies such as enhanced retrieval in Post-Traumatic Stress Disorder (PTSD). While many extant philosophical discussions of memory [11] [63] refer to the case of H.M. who had traumatic damage to his memory system which affected storage, a far more common memory pathology is intrusive memory. Patients with PTSD experience *intrusive memories*. In these memories, a traumatic incident is remembered vividly and often. This is not because the traumatic memory is often relevant, but because something is wrong with the process of retrieval which privileges a certain set of memories (the traumatic ones) without good reason. This may be because of the stored memory itself, or the way in which memory is searched and cued, but most likely involves failures that run through the stages of the retrieval process.

So if retrieval keeps showing up as the key step in memory, we should expect a good

theory of memory function even at an abstract philosophical level to either explain why this is or provide a basis for such an explanation. An objection to this way of thinking about memory function might be that for ideal epistemic agents, retrieval is inessential. Perhaps these agents can access all their stored representations instantly in parallel, or maybe the idea of access as a process itself presumes the system in question is non-ideal.

I concede that under some notion of the ideal agent, this objection is correct. God probably would not need a function to get information from his memory system. So one response is to be more specific about what I mean an ideal agent: a creature that carries out the same operations as we do, but is not limited absolutely by computational difficulty. That is, this kind of agent is capable of doing any algorithmically specified computation, but not, say, processing uncomputable functions. I'm also presuming that we can still talk about efficiency for such an agent. This might seem odd, but it strikes me as an assumption shared by classic Bayesian epistemology, as well as other algorithmic but idealized frameworks. For instance, the Bayesian advocates for conditionalization, but does not take herself to be obligated to rule out the procedure of first transforming the credences linearly, then conditionalizing, then re-transforming. Presumably this is because that would be no better, and much less efficient, than conditionalizing.

In my case, I'm appealing to efficiency to support the necessity of retrieval for the ideal agent in the following way: were the agent not to intelligently store and structure her representations in order to optimize them for retrieval, she might still be able to function at the same level by employing an extremely complicated and computationally costly retrieval mechanism. However, efficiency would best be served by having a retrieval function and representational structure that work together.

One further thing to establish is the criterion for *successful* retrieval. I will assume that this consists in two features: (a) truth/accuracy, and (b) relevance/completeness [1]. This

[1] in computer science, relevance/completeness is referred to as the recall/precision trade-off; these trade-offs both refer to the fact that the more inclusive search results are, the less they can precisely pinpoint connections between the query and the retrieved items

means that the information accessed in response to a query must be on topic, not omit anything of clear importance, and of course, be true or accurate. One thing you might expect to see in addition to (a) and (b) is something like faithfulness, or accuracy *to the past representation*, rather than to a state of affairs in the world. However, faithfulness is not needed to characterize retrieval, and in fact, as I'll discuss later, is often at odds with retrieval.

So I have argued that retrieval is integral to successful epistemic functioning in memory, and that successful retrieval consists in storage and search that makes available true and relevant information. Retrieval is a central norm for memory, in this sense, because it defines a function that differentiates memory success from memory failures across many cases and in many systems; it is not a norm in the sense that any particular memory system or agent necessarily should attempt to satisfy this function, or use this norm as a goal to actively guide processing.

2.2 Sleep and Memory

Animals, including humans, seem to be fairly successful at solving the retrieval problem. We are exposed to an incredible amount of information every day, and yet manage to pull out important facts to highlight and be reminded of later. This suggests that looking at animal memory system might tell us something about how the problem can be solved, and what systems that solve the retrieval problem have in common.

In this section, I'll aim to ground 2 claims: (1) functionally, information stored in animal memory systems tends to acquire "model-like" features over time, and (2) these features are developed, updated and maintained primarily or at least to a significant degree *by the memory system*. I'll examine what it means to be model-like in Section 2.3, but for now, the idea is to let the data guide us toward a sense of what these information changes have in common. (2) is necessary to rule out the reasonable hypothesis that on-line deliberation is doing all the interesting computational work and memory is merely its storage

receptacle.

2.2.1 Predictions

I'll now turn to the connection between sleep and memory in animals. I'll focus on sleep because what's at issue between the various theories is whether the memory system itself performs information transformations. Because we neither deliberate nor perceive during sleep, informational changes that happen in memory during sleep can be attributed more cleanly to the memory system itself as opposed to deliberation or perceptual processing. So sleep takes some possible confounding variables out of the equation.

Every theory on the continuum from preservationism to constructivism can accommodate sleep having some effect on memory. What happens in memory during sleep, however, could be understood under very different hypotheses depending on the background theory:

1. Tagging hypothesis: Deliberative systems tag information changes which are carried out by the memory system (but neither designed or planned by the memory system)

2. Index hypothesis: Memory systems build an index during sleep (to a lesser degree during waking) which is used to pull data from a long term, relatively inert, storage system.

3. Model hypothesis: Memory systems build and alter models during sleep (to a lesser degree during waking) which are used directly in retrieval.

4. Constructive hypothesis: memory is an online process of imagination and so should not be implicated in sleep except in the way that all cognitive operations tend to degrade in conditions of sleep deprivation.

The tagging hypothesis about memory is a way of articulating that memory is informationally inert and therefor fits with preservationism. On the model and index hypotheses, memory performs informational operations, analogous to the kind of curation introduced

in the museum analogy. I won't get into the differences between these two views until the next section; instead, I'll focus on what they have in common. Namely, both predict that animal memory systems, to be successful at solving the retrieval problem, should alter stored representations to make them simpler, more cohesive, and structured according to patterns and generalizations. I'll argue that the data strongly favors both the index and model hypotheses over the tagging and constructive hypotheses. Note that while constructivism and preservationism are two ends of the spectrum, they generate virtually the same prediction here, since both minimize off-line information processing in memory [2].

2.2.2 Data

While there has been a lot of behavioral and higher-level work on long-term changes to memories (most famously by Elizabeth Loftus [57]) and other phenomena which might support my view, in this section I'll concentrate on low-level neuroscientific data[3]. Additionally, the model I'm presenting is meant to be a reinterpretation of a swath of research that investigates the influence of environment and more generally prior beliefs on memory [4]. These models often either give a fully general, domain-neutral theory [3] [46], or focus on small-scale, particular schemas [90]. My project is to give a redescription under which these results are picking out pieces of the same underlying phenomenon.

At a minimum, it's clear that in humans, rodents and drosophila, sleep is critical for memory consolidation. The task of this section is to argue that what happens during sleep is likely to be a process of organization and prioritization. If animal memory is a model

[2]This is one reason I do not consider constructivism in depth in this paper. However, more importantly, constructivism is a theory of episodic memory and this paper discusses memory at a higher degree of generality. As such, the overall memory system might be a model system, and there could be some element of scene construction using that system that we classify as episodic memory. Or put another way: constructivism [65] presupposes the existence of a separate storage (and by my lights, retrieval) system that supplies imaginative construction with its content.

[3]This is in part because these kind of structural question can be answered most directly by looking at mechanisms, but also because the findings I discuss in this section have gotten less attention in the context of these abstract debates about memory function

[4]It's worth noting that some models - such as Zhang and Luck [98] - explain data of the kind I'll discuss without any direct influence of priors on memory (in their case, by a combination of straight remembering and on-line guessing). However, this is plausibly due to differences between working and long-term memory.

system, then these are precisely the features we should expect. On the contrary, other views do not predict this. That suggests an explanatory advantage in favor of the model view.

Several patterns have been well established in the research on the link between human and rodent memory operations and sleep, including the following: enhancement for weak rather than strong memories [32], preferential enhancement of goal-related memories [94][32], and enhancement of temporal ordering information [35][44][32]. All of these facets indicate that informational transformations are occurring. Another line of evidence against the tagging view is the two-way connection between memory enhancement and sleep - in studies in drosophila, researchers found that increasing the memory demand in the environment increased time spent asleep, directly stimulating brain regions responsible for memory increased sleep [33], and enhancing sleep chemically reduced memory deficits [34].These interventional studies indicate that at least in flies, there is a genuine causal connection between memory and sleep.

To look at one finding in some more depth, Drosopoulos et al [35] looked at consolidation of temporally ordered information in humans in a sleep and control condition. They used repeat sequences to separate temporal ordering effects from more general ordering effects. Consistent with other work [32] they found sleep only strengthened the forward temporal sequence and not the backwards one - that is, subjects who had slept between training and testing were no better than those who had not at the task of naming which cue came before some cue that was shown to them. However, they did significantly better than the non-sleeping group at naming which cue came after a cue that was presented to them. A second finding was that weak associations were strengthened more than stronger ones - if an associated was less reliable during the training session, it was more likely to be improved in the testing session.

This and other behavioral effects of sleep have increasingly clear neural correlates, at least in rodent models. For instance, Bendor et al[10] taught rats an association between a tone (left or right), and the location of food (at the left or right end of a track respec-

tively). Following training, the authors recorded from hippocampal neurons. Critically, these neurons represent a spatial map of the environment, and go through 'replay' during slow wave sleep, in which patterns of activation experienced during waking are repeated multiple times in the same order during sleep. They played three sounds (the left and right tones used in the training phase, and an unconditioned control tone) during the two sleep phases as well as waking. Playing the left tone during REM sleep (but not slow wave sleep or waking) enhanced hippocampal replay in the left region of the place field, i.e. the part of the neuronal ensemble that represents the left half of the environment. The same was found for the right tone. Further, playing either tone lead to behavioral changes; the rats were more likely to make errors in favor of the side on which the tone was played during REM sleep. We have copious evidence of a correlation between hippocampal consolidation during sleep and changes in waking behavior, but this experiment suggests a clear causal connection; by biasing signals during REM sleep, the experimenters biased the neural patterns at consolidation, as well as the behaviors during subsequent testing. That is, not only does hippocampal consolidation parallel the information transformations I've been talking about, but it's likely that changes in consolidation *drive* changes in behaviors.

These two lines of evidence (informational effects and causal connection) count against the tagging view in the following way. Since the tagging view says that activity during sleep is merely carrying out the instructions generated during waking, it should predict that stronger associations have an equal or greater sleep boost than weak ones. The causal studies indicate that there is a two-way relationship at play, since enhancing sleep enhances memory and learning.

The model view predicts a computationally taxing and critical period of consolidation, since models have to be assembled and designed in a non-trivial way. One line of evidence in support of this claim is sleep-linked memory triage. This is a broad category that encompasses many ways in which information acquired during the day is altered and consolidated during sleep. In selective item consolidation, subjects tend to forget items which don't fit

into a pattern (for fMRI corroboration on this point see [76]), and reinforce those that do. In multi-item integration, a new stimulus that fits into an existing pattern will be integrated with that pattern, and in gist extraction, a pattern will be generalized leading to confabulating items which fit in with the pattern but were not presented. To simplify a little, it seems like consolidation during sleep is important for fitting new memories into existing patterns, and extracting new patterns.

The phenomena of information transformations like gist extraction is compatible with a wide range of memory structures, though the idea of active forgetting [76] makes some trouble for a strict preservationist picture. What I take to support the model view is that this consolidation takes place during sleep. If reasoning or other on-line processes were responsible for gist extraction and generalization, we would expect items that didn't fit into a pattern to be recalled less even without a period of sleep before testing. However, there is a significant sleep-dependent boost for each of the three effects I've discussed [94]. One explanation for this draws on the neurochemical environment during REM and transition-to-REM (TR) sleep. Poe et al note that the low acetylcholine and high norepinephrine during these phases facilitates long term depotentiation, the reversal or dissolution of long term potentiation, which is often described as the neural correlate of associative strengthening [69].

Sleep-linked processing is by definition off-line, and the neural correlates of these consolidation processes indicate areas associated with memory, rather than other reasoning or learning processes [69]. So it seems likely that these operations are operations of the memory system itself. First, this evidence is predicted by the model view, since consolidation will involve integrating new evidence into existing representations, which has the automatic consequence of generating inferences along the lines of the confabulation cases, where the subject claims to have seen an unseen stimulus 'EFG', which follows the 3-letter pattern of the stimuli which they actually saw.

Stepping back from the details of memory consolidation, another line of evidence ap-

peals to the division of the memory system into episodic and semantic components. Semantic and episodic remembering are two ways of representing which can share some representational content (for instance, I can semantically remember that Fatima was the daughter of Mohammad, or I could episodically remember my teacher telling me that Fatima was his daughter). Whether I prefer one form over the other depends on what features of the information I deem relevant. If I doubt that my teacher is a reliable source, I might want to remember that it was him who told me that fact just in case I get more evidence about his reliability. In other words, the choice of format, just as with models more generally, is relevant for determining how that information will change over time. The challenge for the model view is showing that these large scale patterns are replicated at a lower level. [5]

What does all this data tell us? It paints a picture on which after sleep, our memories are more organized, combined and structured according to patterns, and sometimes expanded and elaborated in accordance with these patterns. This change is not just observed behaviorally, but we can follow the neural correlates of the representations as they change, and make interventions to predictably effect the outcome of this process. These changes are at odds with the tagging and constructive hypotheses, and thus in tension with preservation and constructivism.

2.3 Two Retrieval Systems

What does it mean to structure information according to patterns? In this section, I'll discuss what two conceptual views of memory that capture the kind of information changes observed during sleep: model systems and index systems. For each, I'll define it, relate it to the retrieval norm, and then give some examples of that system in the world.

[5] Buzsaki and Moser suggest that the semantic/episodic divide might be grounded evolutionarily in the allocentric/egocentric distinction in spatial navigation [18]. If this is right, since spatial representations are quite obviously model-like, there's a good reason to think the memory system which evolved from them shares this property. See [28] for evidence that spatial representations are relational.

2.3.1 The Model System

A model system is any information processing system that builds, evaluates, and otherwise processes models. This isn't very helpful - the real question is *what is a model?*

Informally, I'll take models to be representations that work via mapping dimensions of the system being represented (which I'll call the object) onto dimensions of the representational domain (which I'll call the medium). These dimensions can't just be points; what differentiates a model from, say, a group of sentences or a table of facts is that it encodes not just the current state of the object, but also regularities, laws or expectations about how features of the object might change (i.e. dynamics). So for instance, a plastic model of an atom counts as a model because the way that the plastic beads move around on the rings encodes dynamics of the movement of electrons across the levels. Now, note that some features and dimensions of the plastic atom model don't encode anything about atoms; the fact that the rings are painted in colors from blue to red, for instance, says nothing about any property of atoms. So a second property of models is that they are simplified: they function to encode only some of the total features of the object. And further, the model atom represents the dynamics of the electrons via the relationship between its plastic parts - the final property which I'll ascribe to models is that they are relational. If we had a model of the atom and a plastic model of a zoo sitting next to one another, their union does not form a new model. And likewise, within the model atom, were we to separate the various parts based on which particle they represent, this breaking down of the model would result in information loss. This is what makes the atom model a model, but the union of atom and zoo models not a model.

One caveat: I'm not trying to give necessary and/or sufficient conditions for the commonsense category of models. Instead, I'm just presenting criteria for being a certain kind of model that will then explain why models are suited to the retrieval task.

As a class, models are representations which have these properties:

domain-specific: the representation employs structures or uses a medium which

is suitable specifically for the kind of thing it is used to represent, as opposed to being generically suitable for representing anything.

relational: all ways of dividing the representation into an unordered set of local representations will result in a loss of information. This is because some information adheres in the relations between parts of the representation, and so is lost when we separate it into parts.

simplified: the representation has fewer dimensions than there are in the information representable about the set of objects it depicts. (Here dimensionality is understood in the information-theoretic sense to mean a feature space: for each feature or aspect of the data, we add one dimension to the model.)

What is the difference between a model and a network? In the case of a relational database, for instance, we talk about networks having some of the features I've ascribed to models. I take the difference to be that models are *manners of representing*, whereas networks are ways of implementing or storing representations.

Chalmers [22] describes manners of representing as the impure content, where the propositional content that the cat is on the table can be represented in a visual manner, or a conceptual manner and so on. He describes these manners as occurring at a variety of levels of descriptions, such that one representation will usually be characterized by several manners of representation.

This does not that manners of representation are informationally neutral, that is, they sometimes constrain the content *in a context*. Imagine we have an auditory and visual representation which have as their content the same set of propositions about the cat on the table. Now, if we add the proposition that the table is on top of the rug, the two representations may update in divergent ways, adding different entailments to their content. For instance, in the visual representation, it will surely follow that the cat is above the rug. This might not be part of the content in the auditory representation. So while different manners

of representing can have the same content, they won't in general have the same dynamics, or changes in content. Ways of implementing representations, in contrast, are usually informationally neutral.[6]

So the claim that models are ways or manners of representing entails saying that models encode dynamic inference patterns. When the same static content is represented in two different maps, the total content in each case will tend to change in divergent ways when updating on the same evidence.

Putting all this together, what it means to say that memory systems are model systems is that they employ simplified representations with a domain-specific functional architecture. How many of these representation structures there are and how they fit together will depend on the agent or system in question, but the common thread which I'll now argue for depends on the aforementioned characteristics of models along with the idea of using multiple representations in parallel.

Now that we have a working theory of a model system, we can ask: why single out this system as a solution to the retrieval problem? Here's the idea: in order to be effective under conditions (a)-(c), retrieval systems need to be able to deal with an influx of new information that bears on previously learned information. To do so, they need to be **dynamic** - capable of using this new information to update and change the current representations. Second, the system needs to be able to extend current information to new cases - some queries will require extrapolation or interpolation, rather than just reading off what is already recorded. I'll call this ability to extend to new cases **prediction**, though it need not be about the future. Third, the representations in the system need to be **searchable**. These are meant to be central attributes of successful retrieval systems, but not exhaustive.

The relational features of model systems allow for dynamics, since isomorphic ranges essentially encode a series of behaviors, and given that the medium itself is bound by rules

[6]Of course, just like any feature of the implementation of some content, it can constrain the content; the machine might not be big enough for some degree of information, for instance. But in general, two identical representations can be stored in totally different ways without this effecting their dynamics.

(say, spatial rules), the rules on the medium will generate rules for the representation. For instance, take a model of an argument in an a conceptual map on a spatial medium. The isomorphism here is between conceptual nearness and (Euclidean) spatial contiguity. So it follows that the conceptual nearness, according to our model, will always be symmetric. Intending to represent the concept of 'art' as closer to the concept of 'dreams' will result in the symmetric revision that 'dreams' will be closer to 'art', and most likely many more changes.

The simplified nature of models enhances prediction. This is because simplicity is known to increase predictive power by pushing the model away from overfitting. Here's Hitchcock and Sober on this point: "For the complex model to have the higher estimated predictive accuracy, it isn't enough that it merely fit the data better than its simpler competitors (that's pretty much inevitable); it must fit the data sufficiently better to compensate for the loss in simplicity it represents" [47]. And it's easy to see why simplicity enhances search.

Finally, the domain-specific nature of models supports each of the characteristics. Domain-specific representations can encode more regularities, since many regularities hold only within domains and not in general.

The reason to have multiple overlapping models can be derived from the second part of the environmental condition - that the agent expects to get a lot more information in the future. With more and more information, the models the agent currently has might become quickly inaccurate. One reason to have multiple models is to be risk-averse. Part of the distinctive role of memory is to provide a database of facts that can be re-purposed even if the agent's fundamental assumptions and beliefs have changed. Multiple overlapping representations help fulfill this role by allowing the agent to have a range of representations with different dynamics. Episodic memory, for instance, seems like it's slower to change than semantic memory; this fact may account for why folk psychology seems not to treat episodic memories as completely static. So this allows the agent to use episodic memory

as a source of recovery if their semantic memory needs to be radically revised - this means the model system retains prediction, dynamics and effective search in cases of recovery or goal drift.

Overlapping representations allow the agent to use multiple functions; for instance, if I'm having trouble remembering in general what kinds of things people give as housewarming gifts (a semantic task), I can switch to episodic retrieval and think up examples of things people have given me as housewarming gifts. This is particularly true if the search functions have exponential cost and are reasonably independent; in that case, adding a new function will massively increase the likelihood of retrieval. In addition, different manners of representation allow for sequential versus random access retrieval, so the agent can use a flexible retrieval strategy which exploits the advantages of both methods.

2.3.2 The Index System

In the model system, the entirety of stored information is transformed and used in models. The index system, on the contrary, keeps the underlying information in unstructured form, while using a dynamic index to alter what is retrieved over time. As I'll discuss in Section 2.4, these two systems exist on a continuum. While I haven't found anyone directly defending the index system as a model of memory in the philosophical literature, I think it captures an intuitive idea: why can't we just work with unstructured stored data by means of a very intelligent search system? I'll build this thought into a competitor for the model system in this section.

An index system has two representational components: a (mostly) unstructured dataset, and an index. Processing in this system involves constantly updating the index, and retrieving items from the dataset via the index. Essentially, the index functions as an intermediary between the query and the dataset. The index itself can be relatively simple, or very complicated, consisting in multiple ways of retrieving the data and highly structured clusters or properties.

While an index in the back of a book is usually an alphabetical list of terms with a link to a page number in the text, indexes for more complex retrieval systems tend to be *inverted*. This means that instead of a list of terms, we have a list of documents. Each document has a set of entries associated with its content. Unlike a table of contents or index in the back of a book, an inverted index represents the documents through their contents rather than representing some content through its location in documents.

A simple method for indexing is to represent documents as vectors, where each cell records the place of the term in the string. For instance, the sentence "I should store this now" can be represented like this:

$$\begin{bmatrix} ID & should & could & store & this & now & I \\ 1 & 2 & 0 & 3 & 4 & 5 & 1 \end{bmatrix}$$

In this case, I recorded all of the data in the string, but often, I only want to index some of the data: for instance, I could index all of the titles of articles in a news archive[7]. Typically, because even an index this light can become expensive, we would try to reduce the dimensionality in various lossy or lossless ways depending on the application [8]. Then, I could use the simple representation above to look for articles that use the phrase "should store" by looking for numerals n, $n+1$ in the second and fourth columns. For every time that pattern appeared, I would use the entry in the first column to bring up the document containing that term in the headline.

This kind of system is widely used in web search. I'll focus on index systems for web search over natural language data; similar systems are used for searching and indexing images, but natural language data is complicated enough to see the features of indexing that are important for my purposes.

Essentially, all web search operates via indexing; users put data on the web in formats that range from semi-structured (such as a wikipedia entry) to unstructured (such as an

[7] data structures such as tries [39] that store the entire dataset are sometime classified along with indexes, but for my purposes an index must be an intermediary between the query and a further dataset.

[8] for instance, using singular vector decomposition (SVD)

image), and search engines crawl through that content and index it. Cutting edge indexing procedures for engines like Google involve combinations of hand-coding, intelligent use of user input, machine learning and so on. User input can take the form of explicit feedback but more often involves things like taking the fact that they went several pages into the results before clicking as a failure (see [56] and [50] for recent work on making the ranking process more intelligent).

The index part of the index system shares some of the features of a model system, even though the underlying unstructured data is different. The index is usually simplified - unlike the index used above, all of the information in a document is almost never represented. Indices are often relational - an easy way to compress an index is to omit strings shared by all documents, which has the result that the representation of any particular document is determined in part by what other documents are being stored. However, most indices are not domain specific. For instance, an index used for web search handles all kinds of natural language information. In part, this must reflect the practical use we put indices to. However, a genuine difference with models is that an index can be highly content neutral, such as the example above, or very content specific.

2.4 Comparing the Index and Model Systems

How does the index system solve the retrieval problem? It updates the index frequently and intelligently, in a way that involves not only adding entries, but changing up the structure. This allows for effective search by keeping the index sufficiently simple and making sure it represents or draws on relevant features of the data. For example, the index to a database of news articles should probably encode (among other things) dates of those articles. But even better would be an index that had a hierarchical structure, for instance, it could embed an index that separated articles by sentiment under one that characterized them as movie reviews or sports articles. As it turns out, many indexes are organized this way - not unlike tags and sub-tags on a calendar app.

The step I made in the last couple sentences of the proceeding paragraph is a blueprint for the argument of this section. As we ask indexes to do more and more difficult tasks, and grant that they have the necessary computational capacities, an index system as a whole starts to function more and more like a model system.

What makes web search different from human memory? One difference is that while the data in web search is added by users who all have different goals, in our memory systems, the same agent adds data and categorizes it. Even if you take a very modular approach to the mind, there is still both a set of interests that characterize the agent as a whole, and a good deal of calibration between cognitive faculties - even if this calibration is unsupervised and unconscious. In web search, content is often added in an adversarial way. For instance, most companies who have web pages want them to appear as high on the ranking for any search query as possible, whereas the search engine wants only results that are relevant to users (or at least, that users will click on).

This difference explains why web search must use indexes: the task of structuring the data itself, rather than superimposing an index with structure, is much harder than for a human brain with a cooperative and calibrated input system. But even in current search engines, you can see the switch from one system to another. Think about how many times you put in a search query and found your answer just by looking at the page, without clicking on any of the retrieved items. This is a case where the function of retrieval is being satisfied just by the index representation, instead of the index acting as an intermediary. How can the index itself fulfill this function? By being representational, relational, simple, and so on.

An interesting case here is the use of features of the data, rather than the data itself, to answer queries. Xia et al[96] present a model on which the mean and variance of the data is used to update the index instead of the values in the data itself. This is meant for cases with lots of fluctuations in the data around a fairly stable mean; the computational challenges of indexing on-line around this kind of data are significant, so using this higher

order feature instead decreases costs. For instance, location-based services like smartphone navigation involve constant influx of data - Xia et al's system uses the mean and variance of this information for updating the index. Essentially, the index is working like a model.

So I think there's reason to believe that as we relax the computational limitation on an index system like web search, and allow it the power to incorporate more of the content into the index in a hierarchical format, queries and other user demands will be satisfied more and more by the index itself. This can happen even without relaxation, as the Xia et al. case shows - there we just adjusted the relative complexity of the various computational stages. Of course, I'm not suggesting that it's optimal to discard the data. I would agree with most epistemologists that one should never give up information if it doesn't cost you anything. But the data will become less and less relevant to the day-to-day operations of the system. Thus, functionally speaking, as we approach negligible computational limitations, the index system will function just like the model system. If you're not particularly interested in the ideal agent, it's also worth noting that the relevant limitations in the connection between data and its analysis are not particularly applicable in the case of memory in a biological organism.

On a more empirical note, the mammalian memory systems described in Section 2.2 resemble a model system more closely than an index system. One relevant finding is confabulation. The increase in confabulation following sleep, such as the memory that I saw 'EFG' when in fact I saw a set of other alphabetical strings in the same pattern, is not accommodated well by the index view. As I've noted, generalization and gist extract during sleep leads to the creation of new, confabulated memory entries. Confabulated memory entries are at least sometimes produced by the memory system. If memory is an index system, the index is not responsible for generating content. So on the index view, sleep-linked confabulation must either be explained away, or a new mechanism posited to explain the addition of content to the index and the storage system respectively. On the contrary, on the model view, the confabulation results are expected as a consequence of simplification.

For instance, a natural simplification of the sets of letters is something like 'I saw a bunch of strings of 3 consecutive letters' - and of course, the confabulated string fits just as well under this heading.

2.5 Epistemological Context

Here I'll survey some views on memory in the epistemology literature that contrast with the model view. I'll assume the framing of the retrieval problem, and the environmental conditions (a)-(c), and argue that these other options fail under those assumptions to give a satisfactory account of the contribution of memory to knowledge.

2.5.1 Causal Theories

The first class of extant views are causal and preservationist theories. While they have key differences, all these views subscribe to the following:

> **Origination Principle:** For every memory with content p, there is a unique original mental act with content $p*$ to which that memory can be linked (which forms the grounds for identifying its justification and/or content).

This will be the target of my arguments, but first I'll discuss the views themselves.

For Bernecker, who starts with a conceptual analysis of the individual states of memory, memories are mental representations which share a continuous internal causal connection with a sequence of states of the same content, tracing back to a single incident of perception or some other kind of doxastic act. In the latter case, for instance, my memory of the concluding step of a theorem will relate back to my realization of that step. In general, the original incident will be a case of coming to know, coming to believe, or coming to acquire the information which I'll call the content of that incident. When the original incident is a propositional attitude, the content is simply the content of that attitude; if the incident is non-propositional, which will be the case under some views of perception, the content

is something like the proposition that corresponds to the non-propositional content. For the sake of simplicity, I'll focus on the case where the original incident is a propositional attitude.

It's this original incident that gives the memory its content: "memory contents are fixed by the past environments and remain unchanged until some later moment of recollection" (169) [11]. On Bernecker's view, when I'm slow-switched from Earth to twin-Earth, my memories with the content 'water' will refer to water if the original incident was pre-switch, and twin water if it was after the switch. This view falls under a more general class of causal views, where what makes some state a memory depends on a causal connection between that state and the other states which make up the memory trace. Causal theorists are not all committed to sameness of content among the states in the trace; Bernecker allows for entailments of the original content to be added to the later content, for instance, and Michaelian (in earlier work, before rejecting this view to turn to constructivism [64] [63]) allows for change and reconstruction so long as there's a "sufficient" degree of similarity between all the contents in the trace. But in either case, these views all clearly presuppose the Origination Principle.

Causal views of memory have a few advantages. I take it that they're quite intuitive, and express something like our common-sense notion that memories are preserved contents from the past. Causal views also allow for a neat and simple treatment of memory content; since each memory can be traced back to an initiating epistemic act (for instance, my memory of seeing an apple traces to the act of me seeing the apple), it simply inherits the content of that act. Thus, memories refer to the past and restrictions on perception, imagination or other doxastic states which might be initiating events carry forward to restrictions on memory. If I'm not in the position to form beliefs about the paradoxical circle-square, for instance, I won't be able to remember anything about it either[9].

[9]John Campbell[20] offers an alternative to the causal view. He argues that since many memories can't be traced back to an initiating event, it's more plausible that memories are about possible past instances than actual ones. For instance, when I remember my grandfather's gas station, it seems implausible that this entails that I'm remembering a single instance of being at the gas station. Instead, I could have taken elements from

These causal views are a natural fit with a position about the function of memory: preservationism. As presented in Burge [17], an agent is entitled (or prima facie justified) in holding a remembered belief if the belief represented in the memory was originally justified. More generally, if the original incident was of epistemic value (if it was a case of knowledge, justification, responsibleness etc), memory functions to preserve that status. One could extend this view to infer a norm for the memory system, in the following way. First, epistemically well-functioning memory preserves justification. Since justification is inherited from the original belief formation, memory ought to hold on to beliefs in as close to the original condition as possible. It's up to the belief formation and revision system, in other words, to come up with justified beliefs, and the memory system has the more passive role of preserving that justification.

In response to objections that preservationism is a bad fit with psychological evidence, Michaelian proposes a weakening of the thesis on which "for any finite cognizer, a certain pattern of forgetting is necessary if her memory is to perform its function well... by eliminating "clutter" from her memory store, this pattern of forgetting improves the overall shape of the subject's total doxastic state" ([63], 399). On his picture, the ideal agent would preserve all of their memories, but for agents with limited capacities for storage, the norm is to store up to the point when storage inhibits retrieval: "Even where the subject actively updates her beliefs, the record that underwrote an outdated belief can continue to lower the reliability of retrieval" (412).

In summary, the views outlined above are in agreement about several points. First, they define the content of memory or the norms governing memory by reference to a single original incident which is not itself a memory. This requires the Origination Principle. For Burge, this original incident is the original belief (or perceptual belief etc.) which is either

many visits and compiled them into a single representation. Campbell proposes that this representation be treated as a representation of a possible way the gas station could have looked to me. This account avoids the objection to the causal theory that our memories rarely seem to meet causal criteria, but it's vulnerable to an objection about the accuracy of memories. What does it mean to truthfully depict a possible past incident? I'll set Campbell's view aside for now since I can imagine extensions of it that would be consistent with my conclusions, and others that wouldn't be.

justified or not. For causal theorists like Bernecker and Michaelian, it's the incident which fixes the content of the belief. It's natural, but not strictly logically necessary, for these incidents to be the same. A second similarity is that memory is conservative in nature, functioning to maintain or sometimes repair content or justification; while some of these views allow for changes in content, these changes are restricted by the original incident and there is no positive norm provided for memory change. Note that this conservatism depends on the first: without a fixed idea of the original thing, there's nothing to conserve.

My first objection to the origination principle is just that it's unmotivated - what's so special about the original content? A reply might be that we have direct contact with the original contents, or they have a greater degree of justification. There might need to be a system to maintain inferences off of these core contents, but that wouldn't be the job of the memory system.

Note that because original incidents can be perceptions, realizations, emotions and all kinds of other mental acts, there will need to be a privileged content in every case - and to avoid a heterogenous and disjointed system, these contents will have to either be the same or easily translatable between each category. If we look closely at what makes up a remembered episode, I doubt that there will be a natural unit which forms the secure content; instead, while some of the information will be more or less noisy, there won't be a difference in kind. And in fact, sometimes we might use many memories in the origination-principle sense to generate a extrapolated content of which we are more certain than any of the memories. I've argued throughout this paper that we should expect this to be part of ideal memory functioning - for the system itself to generate new content via abstraction, simplification and so on - and yet this content would not count as remembered given the origination principle. Here, Michaelian[63] would object that causal views can still account for some content change and addition. But this is only in non-ideal agents.

Overall, the issue with the origination principle is that it imposes a means of individuating content on the memory system. On the contrary, the considerations I've offered support

an epistemic advantage for the memory system *itself* being responsible for building and changing representations. Memory faces different challenges than other systems, and has unique data needs. Even if these new structures can be described in terms of inherited content, as Michaelian wishes to do, doing so adds nothing to the epistemological theory. Insofar as epistemology and theories of content should go together, origination theories can't work.

2.5.2 Memory-as-Testimony

This section is concerned with a view that doesn't rely at all on the origination principle. In contrast, it treats the epistemic value of content supplied by memory as equivalent to testimony. Given the wide variety of views on testimony, this theory of memory could work in a wide variety of ways. For simplicity, I'll assume the following view of testimony, which Barnett [7] calls the naive view: "when a source of testimony tells you that p, what you learn first is not p itself but instead merely the fact that the source says that p". Barnett goes on to qualify that this is an internalist view in that it depends on your justification in believing that your source is reliable, rather than the source actually being reliable, but I'll stay neutral on this point and allow either version to count as the memory-as-testimony view.

One relevant issue for the testimony view is that the retrieval problem that I've argued is central to memory is outside of the normative scope of the view. Why? Because, just like in the case of ordinary testimony, when we are asking the up-taker of the testimony to justify her reliance on it, this view is about how and when the rememberer is justified in relying on the information served up by her memory system. The retrieval problem is about how to store and modify information in order to offer the correct entries up in response to a query; that is, in the analogy, good retrieval is like being a good testifier, not a good up-taker of testimony.

Of course, it's coherent to ask what a good testifier looks like. I'd like to note here that

answering this question is not part of the analysis offered by extant memory-as-testimony accounts. More problematically, it's the focus of almost none of the testimony literature itself. And in fact, this may be because epistemology should be relatively permissive about what constitutes a good testifier. It doesn't seem irrational for me not to fully optimize my communication to serve your epistemic goals, let alone to prepare my storage systems for years in advance in order to offer up information which will be both accurate and also relevant to your needs. Even if this would be excellent social behavior, it seems excessive to suggest that it would be rationally required. And yet, this is exactly what we would have to do to maintain both the testimony analogy and the importance of the retrieval problem. In conclusion, I have not argued that there is no useful analogy between memory and testimony, but merely that the epistemic analogy between memory and testimony seems unable to explain anything about how to solve the retrieval problem.

2.6 Wrong Level of Abstraction?

You might think everything I've been saying is sort of reasonable for structuring a machine, but not of normative importance. How could something as low-level as manners of representation within the memory system bear on epistemic rationality?

In response, I emphasize that it's not the details of the model system doing the normative work; the details are meant to show how optimizing for one task that looks fairly passive - i.e. retrieval - leads directly to active and generative procedures. That is, the model view of memory says that just by caring about retrieving relevant and accurate information, we end up with a memory system that functions not like a file-box, but like an off-line deliberation system. It moves information around, builds theories, puts things together, and fills in the gaps. The key result isn't how the various stages work, but the fairly abstract fact that achieving the central epistemic goal of memory involves making new content. In a sense, the algorithmic level is just here to demonstrate the relationship between retrieval and active generation. This is where it's especially key that the model

system isn't a system for storing models, but a system for making them.

Put another way, if you are interested in the question of how we are justified in maintaining beliefs over time, you might think that the information-transforming aspects of memory are not relevant to answering this question. While it might be adaptive to come up with new information in memory, it need not effect my justification in believing based on memory in the 'normal' case where the memory is not significantly altered. But I hope to have shown that this analysis which abstracts away from the messy cases is misguided. This approach misses the commonality between the 'normal' case where a memory resembles a past belief closely, and other cases where the memory is accurate but has drifted significantly from sharing a content with any past belief. That is, both cases are the product of computations that enabled those contents to be retrieved, and both reflect background beliefs and values of the agent. Since computations like modeling are the sorts of things we are used to evaluating for rationality and justification, it should be at least an open question whether the etiology behind memory retrieval, even of 'normal', seemingly-unaltered content, is relevant to determining the justification of memorial beliefs.

2.7 Conclusion

A feature of the model system is that it has a significant tendency towards incoherence; since we're building different kinds of models that idealize away from different features on the same data (i.e. they are overlapping), we should expect some of the extensions and extrapolations to contradict one another. Oddly enough, this shows a strength of the model view. That is, allowing memory to itself build and change models gives us a theory on which those changes are necessarily off-line (very off-line, in the case of human memory changes during REM sleep). Incoherence has worried a lot of philosophers, but the model view gives us a case of incoherence strictly speaking, but stripped of its most problematic features. This type of incoherence is between models, rather than within a model, and is processed at a distance from resultant actions (i.e. it's off-line).

A second consequence worth pointing out is that I've talked a lot about how the model system works, but my proposed change to the internal workings of memory also would lead to changes in the external role of memory, by which I mean the way other parts of our cognitive faculties are hooked up to memory. A rough way of characterizing this second shift is to think of how we use a set of facts from a filebox versus how we use a model. Where the former gets inputted and digested, the latter acts more like a guide.

In summary, in this paper, I emphasized the centrality of retrieval to memory, and situated the memory problem in a series of environmental conditions (a)-(c). From this starting point, I considered how memory changes during sleep give us a reason to look for an alternative epistemology of memory that explains why patterns and generalizations over stored information are enhanced and sometimes learned in sleep. It is these structures that explain the more commonly cited cases of constructive memory. I then presented two kinds of system that solve the retrieval problem in part by structuring and generalizing. My preferred option, the model system, makes, updates and maintains a series of overlapping representations that are simplified, domain-specific and relational. I argued that for realistic human creatures as well as ideal, non-limited agents, the model system works the best, and the alternative index system, succeeds insofar as it starts to resemble the model system. In the final sections, I contrasted this essentially productive memory system to the systems implied by two families of theories in the epistemology of memory. This might have convinced you that these other theories need some work, but I was mainly hoping to demonstrate that adopting the model view will result in non-trivial changes to the epistemic role of memory.

I'll leave the reader with this thought: philosophers have long studied on-line, deliberative changes in belief. But it's evident that many significant changes happen off-line, whether transitions from believing P to believing $\neg P$, or realizing that Q is also a possibility, or so on. I suspect that looking closer at the how information changes in memory systems and understanding what makes a successful or unsuccessful change will be a sig-

nificant part of this broader project of the epistemology of off-line belief formation and change.

CHAPTER III

Exploring by Believing

ABSTRACT. Sometimes, we face choices between actions most likely to lead to valuable outcomes, and actions which put us in a better position to learn. These choices exemplify what is called the *exploration/exploitation trade-off*. In computer science and statistics, this trade-off has fruitfully been applied to modulating the way agents or systems make choices over time. In this paper, I argue that the trade-off also extends to belief. We can be torn between two ways of believing, one of which is expected to be more accurate, whereas the other looks like it will lead to more learning opportunity. Further, it is sometimes rationally permissible to choose the latter. I break down the features of action which lead to the trade-off, and then argue for each that it applies equally well to belief. This conclusion is an instance of a systematic, foreseeable way in which what is rational to believe now depends on what one expects to be doing in the future. That is, epistemic rationality fundamentally concerns *time*.

Introduction

In many decision-making scenarios, we can observe a trade-off between choosing the action that maximizes expected reward, or the action most likely to result in learning something new: the **exploration/exploitation trade-off**. For instance, imagine you are choosing between ordering your favorite ice cream flavor or trying a new one. Exploiting consists in picking the option most likely, on your evidence, to have the highest value. Exploring, on the other hand, involves choosing something previously untested or about which you're uncertain. There's a trade-off because the best behavior for exploring (say, trying every flavor once, even banana-tomato) is rarely the behavior that is the most likely to maximize

reward - and vice versa. The striking result, in the case of action, is that these exploratory behaviors that look like seeking out information are rationalized entirely without appealing to the agent valuing information; even if I only love tastiness, I should still sometimes try flavors that seem likely to be disgusting. The task of this paper is to extend the idea of such a trade-off to the case of belief formation and change: should we ever believe solely in order to explore?

Initially, the prospect of a symmetry between exploration in action and exploration in belief might look unlikely. For one, actions are chosen voluntarily, whereas beliefs are formed automatically, without deliberation or an act of the will (see [5] for a discussion of this question). So the exploration/exploitation trade-off might be a decision-theoretic concept that is out of place in the context of belief. Likewise, we usually think of epistemic rationality as universal and unchanging, whereas rational decision-making allows for trade-offs and merely instrumental rational actions.

However, I will argue that there is indeed an exploration/exploitation trade-off in belief, because of the connection between our current beliefs and our dispositions to conduct experiments and explore the space of possibilities. This paper is the first to posit exploration in belief. However, others have argued for exploration in other parts of cognition, for instace Sripada[85] argues for exploration in the act of imagination. While past work on epistemic trade-offs in belief has focused on situationally-driven trade-offs that are arbitrary and often fantastical[43], this paper looks at a learning situation that characterizes most of our epistemic positions in real life, and posits a systematic and easily implemented rule: deviate occasionally from the recommendations of your best policy (i.e. the first-order plan composed of myopically exploitative acts) in the beginning of inquiry.

I concentrate on the beginning of inquiry only because this is a relatively simple and clear application of the trade-off - indeed, at all stages of inquiry and when the stage is indeterminate, other factors such as relative payoffs modulate the trade-off. The beginning of inquiry is determined by how long the inquiry will extend into the future as compared

to how long it has progressed so far, and how much more evidence will be acquired as compared to how much evidence has already been collected, among other things. These features are relational; two agents may have exactly the same evidence about some issue, but if one expects to get more evidence in the future than the other, they may be in different stages of inquiry. This is significant because epistemologists have long assumed (whether implicitly or explicitly) that considerations about what the agent will be doing in the future, and how long they'll have to do it, are irrelevant to epistemic rationality. Consider, for instance, how unlikely it is for a typical case in the literature on peer disagreement to mention what further evidence might be available after the current episode. Along similar lines, convergence arguments (like the one debated in [9] and [48]) ground rational procedures by appealing to the limiting case of infinite evidence. Consequently, one of the goals of this paper is to propose that we attend more seriously to facts about the agent's evidential position over time.

The structure of the paper is as follows: in Section 3.1, I survey the formal literature on the exploration/exploitation trade-off in action and identify some structural features of decision problems which give rise to the trade-off. In Section 3.2, I introduce an example of belief which I'll use to demonstrate what exploration in belief might look like. I then argue that each feature which generates the trade-off in action holds for belief in Section 3.3. The core argument, in Section 3.5.2, appeals to how belief rationally guides and constrains imagination. Section 3.7 discusses objections and the significance of this project and analyzes its relation to other questions in epistemology including doxastic voluntarism and epistemic consequentialism.

3.1 The Exploration/Exploitation Trade-Off

I'll now explain the trade-off through a simplified version of a classic setup in the literature: the multi-armed bandit. I'll start with a brief discussion of the significance of the trade-off. Readers familiar with the trade-off can skip to Section 3.1.3.

3.1.1 Prelude

The ice-cream shop example I've just described illustrates how wanting to get the tastiest treat now and wanting to get information so as to pick the tastiest treat next time are often at odds with one another. This result is significant, because it contradicts a common-sense maxim of decision-making: to do as well as possible at getting good results, pick the options that are most likely to lead to good results. Instead, this trade-off shows that, *predictably*, occasionally choosing options that don't look so good will lead to a better overall outcome[1]

What does this mean for decision theory? Contrast the exploration-exploitation trade-off to Good's theorem[41], the principle that an expected-utility maximizer will never turn down free information. Good's theorem takes a framework for rational action, Savage's decision theory, and shows that optimal agents in this framework always make a certain kind of choice. That is, Good noted that maximizing *expected utility* entails seeking out free information before acting. On the other hand, the exploration-exploitation trade-off shows that agents who chose only to maximize exploitation will tend to get in trouble in many environments: it's is an observation about *actual* utility. The agent who only exploits *expects* her choices to maximize utility, but we know that they likely won't. And note that the trade-off shows that exploitation doesn't pay off across a set of environments that we can specify in advance and which the agent can identify from the inside. You didn't need to know anything hidden about the ice-cream example to see that it makes sense to try a new flavor. Further, as I'll expand on later, the trade-off is also modulated by observable features of the environment; it's better to explore more earlier, and when pay-offs are lower, and exploit more later, and when pay-offs are higher.

So where Good's theorem describes how expected-utility maximers will act, the exploration-exploitation trade-off raises a challenge for the expected-utility maximizer: how can an

[1] see [49] for a formal proof that an agent with a little randomness built in almost always outperforms one that uses a more standard algorithm for approximating rational choice).

agent who always chooses the act that looks best avoid the pitfalls of exploitation? The predictability of the failures of pure exploitation mean that our ideal agent should know better. As I discuss in Section 3.1.3, one possible solution is to maintain that the perfect agent always does know better, and so never has to deviate from the plan. However, this does not make any progress on answering the challenge for the less than perfect agent. The difficulty, and the interest, in understanding the exploration-exploitation trade-off in action is this point: if you know that occasionally deviating from your best plan will get you the best results, does that make these moments of essentially random deviation rational? And what does this imply about what it means to be rational?

3.1.2 The Multi-Armed Bandit

In this example, I show that *as a rule*, in order to receive the optimal reward from many environments, a (somewhat limited) agent should occasionally choose actions not recommended by her best policy. By as a rule, I mean that this result predictably applies to many environments, and that one could reasonably believe that choosing a non-recommended action would help based on limited information. Optimal reward will be measured by aggregate preference satisfaction, which in this toy example will be total number of dollars won. This section serves two functions: (a) it explicates the exploration/exploitation trade-off for action, and (b) it establishes that some behaviors that seem to reflect a preference for information are rationalizable for agents who do not intrinsically value information. That is, if exploring non-recommended options is predictably associated with optimal reward, then rational agents will carry out these behaviors regardless of what they take to be optimal reward.

Now, some setup. Let's assume a simple expected utility (EU) framework. We have some agent, who has probabilities over various outcomes and multiplies these with the corresponding utilities to generate expected utilities. Canonically, these outcomes are complete states of the world. However, in practice, we often idealize away from these complex

states into simpler ones, and evaluate only the value of the immediate result of each action. For now, our expected utility framework will be near-sighted or *myopic* in this way.

Now here's the problem our agent faces. She can choose to play at one of three slot machine arms $i - k$. After each play, she may continue at the same arm, or switch to a different arm - in other words, this is a sequential choice problem. Each arm produces stochastic rewards distributed around a fixed unknown bias [2]. Let's say she starts with the following estimations of the biases, where a higher bias means a higher probability of a valuable outcome: $b_i = .5, b_j = .2, b_k = .1$. Now, assuming that she's going to play these slot machines for some significant amount of time, what should she do?

One method would be to always choose the arm with highest expected reward, calculated from the estimated bias and her confidence in that estimate. She would start by choosing arm i. After she plays i, she'll get some information. Let's say that the true bias of i is .8, and the outcomes in the short term reflect that bias fairly faithfully. So by using this method, she will continue to choose i over and over again, because its estimated payout will never drop below that of arm j, which has the next highest estimate. This is the method recommended by her myopic expected utility rule; I'll call it the myopic exploitation method. It's exploitative because it always does what is best according to current expectations - where A is the act which maximizes expected utility, myopic exploitation always requires her to do A.

How good is the myopic exploitation method? Only as good as our agent's expectations. If she's right initially that i is the best arm, she'll attain the optimal reward. However, if she's wrong, and for instance k actually has a bias of $.9$, her total reward will not be optimal, and indeed will be significantly suboptimal as the choice is repeated over and over. She has no reason to try the other arms if she only acts to maximize reward at the next step. Myopic exploitation has a significant risk of getting her stuck in a local maximum, a section of the reward landscape that is better than all neighboring possibilities but not the

[2]Multi-armed bandit problems tend to have looser assumptions around bias, for instance that the reward state evolves according to some unknown Markovian function [59]

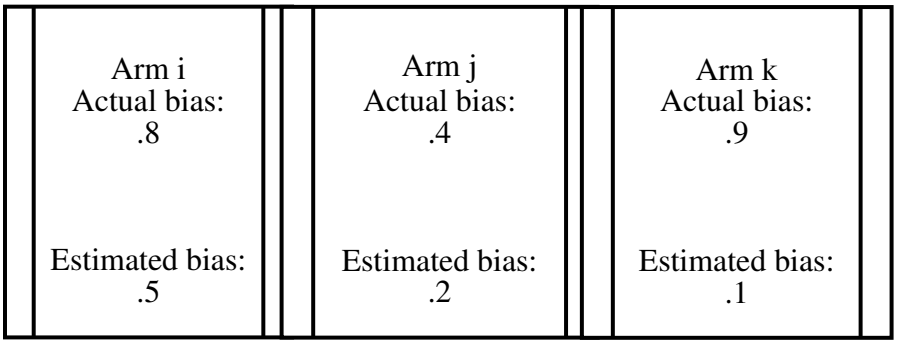

Figure 3.1: A simple bandit problem, where arms have a bias drawn from an underlying distribution and fixed noise

best overall. Once she's in the bandit situation described above, she'll never stop making the same suboptimal choice.

A very simple way of allowing for exploration in an exploitative decision strategy (where A is the act with highest exploitative value) is to add this rule: at every decision point, choose a random act other than A with probability ϵ, choose A with probability $1 - \epsilon$. This is called an ϵ-greedy strategy. Here, as we increase ϵ, our agent explores more and more, and correspondingly, any decrease in ϵ will lead to an increase in exploitation. As ϵ approaches 1, our agent will learn a lot by choosing all the options equally, but her learning will not benefit her, since her knowledge about the options won't affect her behavior at all. As ϵ approaches 0, her behavior will converge to the myopic exploitation rule - she'll always maximize, and never veer off course. Because she will learn more and more about her environment as she makes these choices, it's reasonable for her to start off exploring a lot and then exploit more and more as information accumulates - when she knows everything about the outcomes, there's no need to try new things, whereas when her expectations are poorly informed, maximizing expected utility is unlikely to be particularly effective.

For instance, in the multi-armed bandit case sketched above, an ϵ-greedy strategy with a sufficiently large ϵ will cause our agent to occasionally sample the arms other than i. As she does so, her confidence will rise that arm k is actually better. As we lower ϵ, she'll only

choose k, which is the optimal strategy.

As it happens, ϵ-greedy methods approximate optimal solutions to the multi-armed bandit problems in many contexts. The extant optimal solutions involve calculating the Gittens index of each arm, which is roughly the value that we place on continuing to use that arm adjusting for the potential of learning. This approach splits the high-dimensional optimization problem into a series of smaller problems: calculating the relative values of trying each arm once at each state. [3]

As I've described, the exploration/exploitation trade-off is typically formally specified in the context of the multi-armed bandit problem. However, it is used far more widely to describe a trade-off in decision-making, for instance in clinical trials[70] and to describe foraging behavior in birds[54]. The common element in all of these applications is that the agents face a series of choices that predictably, but with uncertainty, produce both rewards and evidence. Further, there is a systematic link between some actions (foraging at a new patch of ground, giving a patient a medication) and some evidence (learning how many seeds are present in that area, or a new datapoint about how effective the medication is). This connection between act and (expected) evidence is distinct from the connection between action and (expected) reward. That is, the bird might expect that the new area would not be very good for seeds, but also that sampling from it would yield significant information. It is these features that give rise to the trade-off, and make exploration part of the optimal strategy even for agents who do not deem information to be intrinsically valuable.

[3] Solving the problem involves calculating the index for each arm i , given by the following equation:

$$v^i(x^i) = \max_{t>0} \frac{\mathbb{E}[\sum_{t=0}^{\tau} \beta^t r^i(X_t^i)|X_0^i = x^i]}{\mathbb{E}[\sum_{t=0}^{\tau} \beta^t|X_0^i = x^i]}$$

Where τ is a stopping time, r is a reward, $x^i \in X^i$ is a state and β is the survival probability, which is the probability that the situation continues into the next iteration. Then, the optimal policy is to always play the arm with the highest Gittens index at each step. This is a computationally expensive procedure (relative to approximations such as Upper Confidence Bound[6] (UCB) and ϵ-greedy Q-learning) that relies on forward induction [59]. Crucially, the Gittens index of each arm typically declines after each play, so the agent does not continue to play the same arm even if it generates high reward.

3.1.3 Extending the Trade-Off

This section takes the basic framework from the previous section, and draws out some conclusions for practical normativity that I will rely on in the rest of the paper. While the exploration/exploitation trade-off is well studied in other fields, it has received little attention in philosophical rational choice theory.

You might be thinking that this trade-off only occurs because of failures particular to myopic expectations. However, the trade-off also arises for agents who plan many steps ahead. One series of examples comes from reinforcement learning (RL)[87], a framework for learning in Artifical Intelligence that has been used to model animal and human cognition among many other applications. In RL, the agent calculates the value of each progressive step that she might possibly take, multiplied by a discounting factor [4]. Reinforcement learning algorithms can plan over arbitrarily many future steps, and yet standard models include perturbations designed to induce exploration. Exploration is explicitly encoded in a wide range of RL methods, from basic algorithms such as Q-learning to more complex model-based methods. So the exploration/exploitation trade-off is not dependent on myopia.

However, this does not demonstrate that the exploration/explotation trade-off arises independent of computational limitations. For one, the reward system in RL is somewhat different than in classical expected utility theory due to the intractability of calculating every possible state of the world. In particular, rewards in RL are thought of as inputs from an external system (but usually one internal to the agent) - this means that the value associated with learning a piece of information can be manipulated by the agent, unlike the more static, hard-coded use in an orthodox EU calculation. To encode the value of information, RL systems often encode a direct value for novelty or surprise[84][23]. There is still a reason to explore in RL because there's still a risk of getting stuck in a local

[4] this sometimes includes every possible act, or is cut off at a future horizon - see [52] for arguments that employing a horizon may actually be optimal

maximum. The only difference is that now we're concerned with a maximum over whole plans rather than single choices.

Now let's look at what happens to the trade-off in an orthodox decision theory context where choices are over complete states of the world. Can we map exploitation onto maximizing expected value - and if so, what is exploration? It might at first seem like balancing exploration and exploitation would not be a problem for a perfect EU agent. For every exploration-type advantage, there is an expected-utility rationale. For instance, I should try each arm because even if there is a small likelihood of finding one better than j, the cumulative long-term reward of switching in that world is quite high. Also, if expected utility maximization is the paradigm of practical rationality, it shouldn't be this easy to find a rationale for violating it. However, Rothschild[77] proves that there is a positive probability that an optimal, EU-maximizing agent will settle on a policy of choosing the wrong arm and continue that way forever. This proof assumes an algorithmic, forward-looking agent, and that the agent correctly believes that the function generating reward is fixed.

Could a different approach to sequential decision-making articulate an optimal EU solution that would never get stuck infinitely with the wrong strategy? This question goes somewhat outside the scope of the present paper, but I'll note two ways to deal with it.

One way of incorporating the trade-off here is to allow that exploitation value is only a heuristic. In making this concession, the trade-off is under the umbrella of EU theory, but still explains something about how EU-maximizers function. Appealing to the trade-off will help explain things like variations in behavior over time, answering questions such as "why did she choose that lever even though it looks unlikely to pay off?". But since a perfect EU agent will optimize the solution to the multi-armed bandit problem, the trade-off between exploration and exploitation is a feature of the expected utility calculation, not a further factor or contraint. This is the view of the relationship between the trade-off and expected utility maximization that I will assume throughout this paper. However, I'll briefly describe another way of understanding the relationship.

The second option is to insist that while zooming out to this higher-order level can accommodate the behaviors I've described in the multi-armed bandit case, there's a trade-off at work here too. That is, exploration is a way of hedging our expectations, and so it should apply even if these expectations are complex and higher-order. Nearly any optimal EU agent with misleading evidence will benefit from exploring, where that means going off policy. While the simple multi-armed bandit I presented isn't complex enough to display the value of veering off of a full EU strategy, all we need to do is expand it to involve more radical uncertainty (for instance, uncertainty about what the options are, or about where new information might come from). So on this view, exploitation maximizes expected value, exploration puts the agent in a position to learn, and the right combination of the two leads to maximum cumulative actual (not expected) value. So it's somewhat underdetermined whether or not the trade-off holds for expected utility maximizers.

What else can we determine about where the trade-off will be observed? In deciding how to collect food in some landscapes, an animal benefits from deviating from the strategy of choosing the patch that looks likely to contain the most food. In other landscapes, there is no reason to explore. And in still others, there is nothing that could be called exploring at all. So how can we tell the first kind of environment from the others?

There are two ways to argue that the exploration/exploitation trade-off applies in a new case without modeling competing methods directly. First, we might demonstrate that the new case is merely a superficial transformation of an old case. For instance, we might re-describe the case of clinical trials as a multi-armed bandit problem. However, until we can demonstrate better outcomes through the use of an exploratory strategy in the new case, it is in principle possible that any redescription is skipping over relevant differences between the cases. A second strategy is to derive general features that seem to apply to most or all cases in which the trade-off obtains. If these features obtain in the new case, then we have a reason to expect the trade-off to obtain in the new case as well. That is, the first strategy relies on a one-to-one similarity between cases, the second on categorizing the new case

according to features observed across a wide range of past cases.

In pursuit of the second strategy, I'll provide a brief gloss on four features of decision problems that lead to a meaningful application of the trade-off. Considering the foraging environments, any setup with only a handful of chances to collect food will not be solved by exploration, since there won't be enough future chances to put new information to work. The animal must not already have enough information to understand the relevant features of the environment or exploration will not be beneficial; conversely, they must have some information about the environment, or exploitation will not be meaningful. The problem must involve uncertainty, but not be totally blind. The animal's behavior must be systematically linked with acquiring evidence; that is, there must be behaviors that predictably raise the probability of getting new evidence, otherwise exploration would be impossible. More interestingly, these behaviors must not line up perfectly with the behaviors that generate value. It can't be that feeding from each patch is always (predictably) good for getting new evidence in proportion to how (predictably) good it is for getting more food. The problem must be a sequential one, with a sufficient number of iterations. In short, the exploration/exploitation trade-off is meaningful when reward and evidence are both linked to acts (conditions 1 & 2), the degree to which an act generates reward is not directly proportional to the degree to which the same act generates evidence (condition 3), and a decision problem is iterated over and over (condition 4). I will return to these conditions in Section 3.3.

A worry underlying the possibility of extending the trade-off is that this trade-off might be merely a feature of the formalism, absent or not explanatory in the informal context. If you weren't convinced by the ice cream example, consider this lyric from a Frankie Ballard song: 'how am I ever gonna get to be old and wise, if I ain't ever young and crazy?". I think this expresses a common sense idea. When you're young, you have an extra reason to act crazy - or to deviate from the action that looks like the best bet from a strategic perspective. The best action for learning is not always the most subjectively rational. The modulation of

the trade-off over time is what makes it explanatory. Young Frankie Ballard should say yes to things that old Frankie Ballard should not, for the same reasons as in the formal case.

Finally, one feature of the exploration/exploitation trade-off in action will be crucial for belief: the relationship between exploration/exploitation and time. As I noted at the end of Section 3.1.2, there's a somewhat generic rationale for preferring to explore more earlier and exploit more later. This reflects a relationship between time and uncertainty - since exploration is more important when uncertainty is high. However, even while holding uncertainty fixed, there's a relationship with time. The information which is reached by exploring has more value when our agent will have a lot more chances to play the slot machines. As she approaches the end of her interaction with the current environment, the diminishing of future opportunities favors exploitation. This is so even if she is still quite uncertain. Take two agents who are equally uncertain, one pulling the first lever of a long sequence and the other pulling the final lever. The first agent has more reason to explore than the second. Of course, in a real-life situation, the boundaries of one context are not given objectively from the world but the agent herself plays a role in defining what counts as the same problem, and in acting in ways that change how problems extend over time.

Reward changes in the environment also modulate the trade-off. Traca and Rudin [89] show that in environments with periodic reward functions, it's optimal to exploit more during high-reward periods and explore more during low-reward periods. In their case, the varying rewards were due to daily traffic patterns on a website, and at higher traffic times, the recommender algorithm did best by exploiting more, and by exploring more at lower traffic times. In summary, variations in uncertainty, potential for actions, and total available reward all modulate the exploration/exploitation trade-off in action.

3.2 A Case of Exploratory Belief

3.2.1 Why Belief?

I'll now turn to the case of belief[5]. This involves changing our focus from practical value, e.g. attaining dollars from a slot machine, to epistemic value, e.g. acquiring an accurate model of an environment. That is, while it might make sense to value whichever beliefs will make me the most money, I'm interested in the kind of value beliefs have when they are true, regardless of their usefulness.

If there is an exploration-exploitation trade-off in belief, this will have two significant consequences. First, just like exploring by choosing a flavor at random could be rational, believing by adopting a belief at random might be rational. Second, just like the initial bandit case had getting the best pay-off at the current draw in tension with getting the best pay-offs in the long run, believing with the greatest accuracy now might be at odds with getting to the most accurate belief in the long run. And all of this would hold without adding in any non-epistemic value.

3.2.2 The Example

The question I want to ask is: do agents who do well at acquiring epistemically valuable beliefs exhibit the exploration/exploitation trade-off? To answer this question, I'll start with a literary example. My aim is to present this example as an intuitive and phenomenologically realistic case of exploratory belief. However, the more realistic the example, the more complex it tends to be, and so I hope the reader will at this stage accept the messiness of this case as arising from the same features which give it psychological realism.

The description below of an incoherent mental state comes from Meša Selimović's novel *Death and the Dervish*. The context is that the narrator, Sheikh Nurudin, has just

[5] A different version of this paper would target credences (partial beliefs) instead of full belief. This would have the advantage of more precision, but full belief is accepted as the subject of epistemology by a larger set of scholars. While I can't go into details here, the argument for credences would not target probabilistic incoherence but instead incoherence in the representations themselves.

passed up an opportunity to mention his incarcerated brother to a powerful woman who might have been willing to help for the right price. The woman has her own agenda; she had offered to help Nurudin on the condition that he falsely accuse her own brother. But Nurudin refused to cooperate. Nurudin belongs to a Sufi order that holds that the best life is disengaged from politics and invovles no compromises, so the false accusation would be blatantly against his moral principles. After rejecting the chance to free his brother (that is, after doing the right thing according to his Sufi beliefs), Nurudin experiences a feeling of inexplicable anguish:

> *I had done nothing wrong, but my thoughts were assailed by memories of the dead silence, of the impenetrable darkness and strange, glimmering lights; of the ugly tension and the time I had spent anxiously waiting; of our shameful secrets and thoughts disguised by smiles. I felt as if I had missed something, as if I had made a mistake somewhere, although I did not know where, or how. I did not know. But I was not at peace. I could hardly bear this feeling of uneasiness, this anxiety whose source I could not determine. Maybe it was because I had not mentioned my brother, because I had not insisted that we talk about him. But I had done that on purpose, in order not to spoil my chances. Or was it because I had taken part in an shameful conversation and heard shameful intentions without opposing them, without protecting an innocent man? Only I had had my own reasons, which were more important than all of that, and it would not be right for me to reproach myself too much. For each of my actions I found an excuse, yet my distress remained.*

Nurudin is in a conflicted state: he thinks he's done nothing wrong, but in another sense, knows that he has ("I had done nothing wrong" but "I felt as if I had missed something"[6]). Furthermore, while he is in a situation of moral urgency (figuring out what to do about his

[6]here, "I felt" is a translation of "činilo mi se" in the original, which does not connote anything particularly emotional

brother, as well as figuring out if his actions are morally wrong), it is not a conflict between truth and morality. Rather, figuring out the truth - whatever it happens to be - is morally significant. Nurudin is torn between background beliefs that the Sufi system is morally correct, and new evidence which suggests that it may be flawed. In response to this conflict in his evidential situation, Nurudin adopts a divided mental state, where the two conflicting attitudes play somewhat different roles. That is, the thought that he has made a mistake is evidenced by his fervid rumination on the past events, searching for where he went wrong. On the other hand, the thought that he did nothing wrong is evidenced by his conscious belief and is more in line with his previous beliefs and convictions.

> **Conflicted Nurudin**: Nurudin is receiving conflicting evidence about how to live that may contradict his deeply held convictions about a Sufi life of spiritual purity and removal from political affairs. He responds by sometimes believing that God exists, and the Sufi theory of renouncing political involvement is morally correct. At other times, he responds by disbelieving in God and judging the Sufi life to be deluded and morally wrong. These switches are not brought on by changes in evidence, and further he sometimes uses both contradictory beliefs in different ways of thinking at the very same time.

As it turns out in the book, Nurudin is at the beginning of a slow shift from being withdrawn from the world, comfortable in a rigid belief system, to getting involved in an insurrection and filled with doubt about everything. But this drift from one state to the other takes up the whole book (which occurs over something like a year in real time) and in this period, he is deeply incoherent.

I'll model Nurudin as if he flips his belief 'switch' between one context and the next, going back and forth from believing in the moral propositions which make up the Sufi project, to rejecting those same beliefs. In the passage above, he is in two contexts simultaneously, since the two beliefs each play a role in a different kind of cognitive activity. At other times, he may fully believe or disbelieve; maybe he's a true believer when in the

tekija (a monastery-like place where the dervishes live), and full of doubt when moving about the city. But note that the shifts cannot be due to evidence, since he is in two of these states above while presumably having exactly the same evidence (given that he is one person at one time). Further, it's hard ot imagine what kind of evidence would justify repeated, context-driven oscillations - and it's even more of a stretch to stipulate that kind of evidence is present in Nurudin's case without any further information. If this strikes you as unintuitive, imagine Nurudin as someone who is optimistic in the mornings and pessimistic at night, and so believes that God exists in his optimistic mornings, and conversely at nights.

By modeling Nurudin as oscillating between beliefs, I am rejecting two others ways to treat the case: as involving a consistent attitude of suspension of judgment or partial (but coherent) degree of belief. These readings need not be impossible, but I'll note some reasons why I am treating this case differently. Suspension of judgment is a doxastic attitude distinct from belief and disbelief characterized by the agent stopping short of coming to a verdict about the truth of some proposition (the status of suspension as an attitude is somewhat controversial, see [40] for discussion). As far as suspension of judgment, Nurudin is unusually conflicted, and his conflict involves many of the behaviors associated with belief: he asserts that he believes, his behavior is explained by a conviction in a proposition, and so on. Suspension of judgment is an alternative response, and one that I'll compare to his actual response later on.

Having a partial (but coherent) credence in the moral propositions in question might sometimes express itself in belief-like and disbelief-like behaviors depending on the stakes. However, we would expect these behavioral switches to correspond to switches in the stakes and/or odds, and by stipulation, Nurudin is oscillating back and forth by following an internal routine rather than an external shift. This doesn't mean that it is impossible to model him in either of these two alternative ways, but merely that it is plausible to model his case as one of internally-driven, non-evidential oscillation between incompatible beliefs.

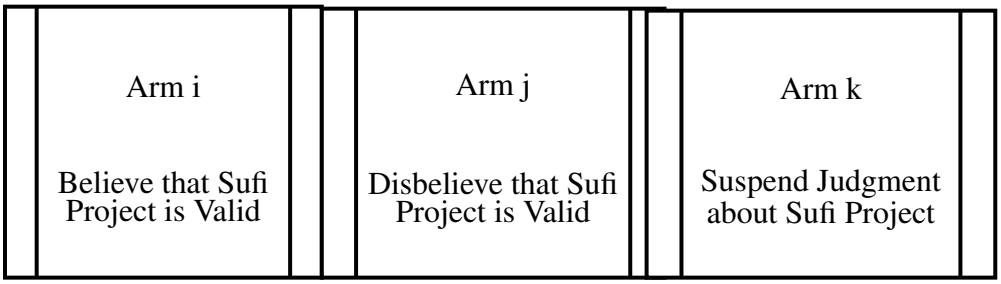

Figure 3.2: An epistemological version of the bandit problem, where choosing an arm represents changing your belief state. Just as the agent in Figure 3.1 rationally alternates between the arms even though endlessly repeating the highest estimated arm has the maximum myopic value, Nurudin rationally samples from all of the arms in this figure, or perhaps alternates between j and k, rather than picking the myopically optimal arm j over and over.

As such, by adopting the strategy I've described, Nurudin is surely taking a hit in myopic epistemic value.

Further, Nurudin presumably does not plan on being in this divided state forever; instead, he's developing two incoherent projects in parallel in order to eventually be able to figure out which is better. Since the belief that the Sufi system is valid is one that comes with an entire moral and descriptive framework, it's reasonable to think that either future coherent state will have very different standards of evaluation, and recommend distinct experiments. So Nurudin is also in a state of meta-uncertainty, which he responds to by moving to a less accurate state (in this case by *all* standards) that might improve his prospect of learning.

3.3 Argument for a Belief/Action Symmetry

In this section, I argue that the multi-armed bandit and the problem Nurudin faces contain essentially the same structure. Therefore, the exploration/exploitation trade-off is operative in both cases. More generally, I present a view on which the trade-off should be a normal part of our reasoning about what to believe.

To make this argument, I'll sketch the features of the bandit case, and then extend them

to Nurudin's. It's important to note here that parity between the bandit and Nurudin is a stronger criterion than necessary to establish the existence of an exploration/exploitation trade-off in belief; multi-armed bandits are one of many problems whose solutions exhibit this trade-off ranging from tree search to applied problems in robotics.

Here's an overview of these conditions:

	action (standard) bandit	*belief bandit*
Condition 1: act generates evidence:	usually	usually
Condition 2: act results in reward:	sometimes	sometimes
Condition 3: evidence and reward come apart:	usually	sometimes
Condition 4: procedure is iterated:	approximately	approximately

As I noted, bandit problems exhibit a trade-off because we expect pulling the lever to give us evidence about the underlying function, and also a reward. These features come apart, as in the case of a high-reward lever that has been pulled many times, so that one more pull will likely provide little evidence but a lot of money. And finally the process needs to be iterated - otherwise exploitation would always trump exploration. I aim to show that all of these features are shared with belief: belief changes what kinds of evidence we can expect to receive based on our dispositions to imagine and conduct experiments, it gives us a 'reward' in the form of accurate beliefs, and these two can come apart in cases like Nurudin's. Finally, each act of forming a belief, like each action, is in some sense perfectly unique. But in both cases, we're engaged in a complex process that can be approximated for some purposes by treating it as a series of iterated moves.

A background issue here concerns internalism and externalism. An externalist version of these conditions would require that, for instance, reward and evidence actually generally come apart, regardless of whether the agent could plausibly be expected to know that fact. On the other hand, a standard internalist version would require that the agent be able to predict reasonably well when the two would come apart in order for it to be rational for

her to respond to this in her beliefs. A pure internalist version might allow that reward and evidence might even be fully coincident in the environment but allow a trade-off so long as the agent reasonably (and falsely) believes that they will come apart. This is a deep issue about the structure of epistemic normativity that I can't hope to adjudicate here. In lieu of that, I'll proceed using the standard internalist version - that is, by arguing that these conditions both hold of the environment and can usually be tracked by agents in an internally predictable fashion. I use this conception because it combines the external and the internal requirements, and so by showing that it can be satisfied, I can also establish that the weaker conditions (pure internalist and externalist) can also be satisfied.

There are some key differences between this 'belief bandit' and the standard multi-armed bandit problem. Most significantly, in Nurudin's case, the value of the various arms are not independent, since they concern belief (or suspension) about a single proposition. In contrast, in the standard bandit, the pay-offs of each arm are independent of one another. This means that in principle, repeated sampling from a single arm in the 'belief bandit' will give you information about the pay-offs of the other arms. However, in practice, the pay-offs of a belief that Sufism is correct concern other propositions as well, for instance beliefs about particular Sufi principles, Quranic interpretation, and so on. And likewise, the belief that Sufism is mistaken will lead Nurudin to learn about other propositions such as the effectiveness of various forms of political action, or the location of a political meeting, and so on (see Figure 3.3 for an illustration). These propositions may not be fully independent of one another, but they stand in complex dependency relations such that an agent with imperfect information or who is not logically omniscientg will never be able to recover the truth of a proposition about the agenda of an undercover political group in Sarajevo from propositions about a Medieval interpretation of the Quran.

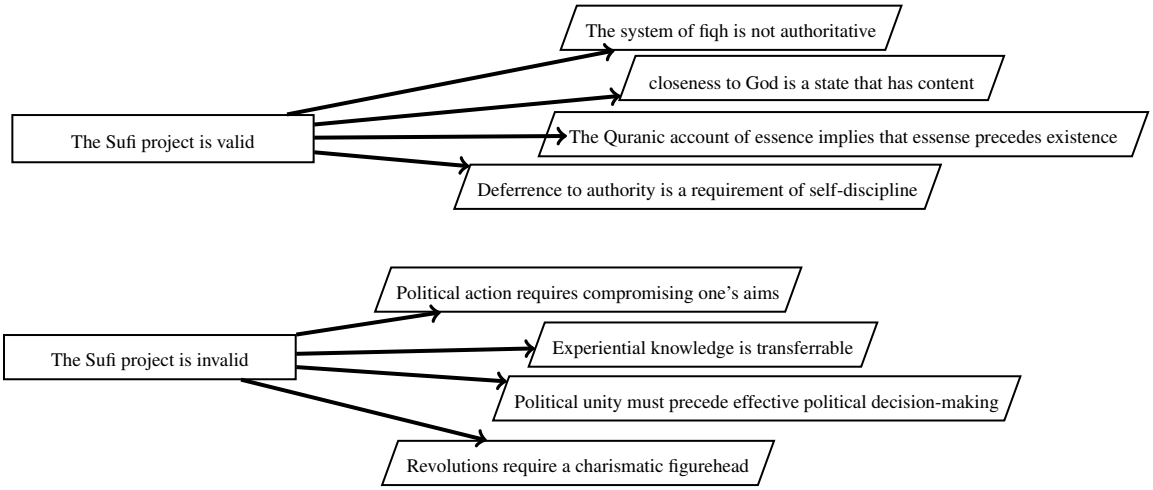

Figure 3.3: Nurudin's position on the proposition in question will lead him to consider and form beliefs about different sets of propositions, whose truth is related in complex ways to the truth of the original proposition.

3.4 Conditions 1 and 2

Conditions 1 and 2 are easily satisfied by belief. For condition 1, beliefs lead to the acquisition of evidence through experimentation and imagination. The most obvious case here is that of methodological beliefs - if you believe that particle collision is not a very good method of discovery, this will lead you to conduct different experiments and so receive different evidence than were you to believe differently. This is not a fluke. Because our intervention in the environment and our process of imagination are guided by our beliefs, they will fluctuate as our beliefs fluctuate.

For condition 2, if we treat accuracy as reward, as is commonly done in any kind of consequentialist epistemology, beliefs are predictably evaluable for reward. (see Section 3.7 for a definition and discussion of epistemic consequentialism). That is, when your belief is accurate, this constitutes a reward. Of course, we don't always know when our beliefs are correct. But in plenty of cases, we find out whether we were correct or incorrect.

Even for the non-consequentialist, beliefs can be treated as generating reward. This just means distinguishing two kinds of intrinsic, non-consequentialist value: the kind of value

beliefs have purely in virtue of their internal components, and some further kind of value. For instance, on many accounts, beliefs have additional epistemic value if they amount to knowledge, and since this value is not perfectly luminous to the agent, we can treat it as a 'result'. In this context, that just means there is some epistemic distance between attaining the state and attaining the reward: you can know that you're in the former state (believing that P) without knowing you've attained the latter (knowing that P). Likewise, it may be the act of eating ice-cream is inseparable from its intrinsic tastiness value, but that fact is sometimes inaccessible to me and so can't play a direct role in planning. I'll return to the question of whether my argument presupposes epistemic consequentialism in section Section 3.7.

3.5 Condition 3

This condition is doing most of the work by pulling current reward and future evidence apart, so I treat it in more depth. I present two connections between belief and other acts that result in a gap between the expected reward (i.e. accuracy or truth) of a package of beliefs, and the expected evidential value (i.e. likelihood to lead to new and/or significant information being gleaned). The first connection is a familiar one between belief and experimentation, and the second is the more rarely discussed connection between belief and imaginative search. In addition to demonstrating that condition 3 obtains, I aim to uncover what features and limitations an agent and her environment must have in order for this condition to obtain.

3.5.1 Belief and Evidence-Gathering

If there is such a thing as rational evidence-gathering, it depends at least indirectly on what you believe, involving choosing interventions that you believe are likely to be effective, depending on background beliefs about how the results come about, and so on. For condition 3 to obtain, we need to sometimes be in a position where one way of believing

is best supported by current evidence, but another way of believing is more likely to result in gaining new evidence, particularly if we are mistaken. Experimentation and other forms of evidence gathering are points at which our beliefs affect our future evidence, so a natural way to describe exploratory beliefs is through this connection. In brief, our beliefs constrain our experimentation, and we can sometimes get stuck choosing the wrong experiments over and over because they are justified by our current evidence. So perturbing our beliefs can solve the problem by forcing us to occassionally choose experiments and gather evidence in ways that are exploratory. Nurudin, for instance, will gather evidence differently in his two ways of believing, and there are reasons to think this is to his benefit; he's able to break out of the restrictive moral framework in part because his seemingly irrational beliefs push him to act in various odd ways that bring new evidence. Likewise, Railton[72] and Kitcher[53] raise the possibility that individual scientists being committed to a hypothesis *beyond* what the evidence supports might help the scientific community arrive at truth in the long run, in part by incentivising the right sort of experiments.

However, maybe all we need is to accept[7] some helpful propositions rather than change our beliefs. On this view, evidence and reward need not come apart, because reward attaches to belief and evidence attaches to acceptance. An intuitive response to Nurudin's case is that the best thing for him to do would be to accept the two incoherent propositions about moral living in different contexts, rather than believe them. Railton's scientist should just accept that his hypothesis is true, rather than believe it. This is presumably what would be recommended by Van Fraassen [91] among others. Note that the disagreement between my view and the acceptance view is about a particular role of acceptance, namely whether exploratory acceptance can explain all epistemically rational exploration; I do not intend to deny that acceptance is a genuine, rational mental state in many cases.

What is meant by acceptance? On one end of the spectrum is the view that acceptance

[7] I here discuss acceptance instead of supposition. While supposing that P and accepting that P are both alternatives to believing that P, they differ in that supposing that P would not lead you to experiment as though P were true. To simplify somewhat, you suppose for the purpose of thought but accept for the purpose of action.

just involves acting as if something were the case without changing any beliefs or making any new inferences. On the other end is the idea that acceptance is descriptively identical to belief, it just falls under different normative standards. The latter in this context is question-begging, since I'm arguing that the normative standards for belief should cover exploration as well. Likewise for the suggestion that acceptance and belief differ only insofar as beliefs are only sensitive to direct, truth-seeking considerations (see Shah and Velleman [82]).

I'll consider two versions of the acceptence objection, one where acceptance is just acting-as-if, and the other where it involves some doxastic changes but not as far as in belief (perhaps inferences drawn from the accepted proposition are quarantined away from the agents other beliefs, for instance).

In the first case, if Nurudin accepts and does not accept the Sufi moral system in different contexts, but suspends judgment, there are two issues. First, it's not clear that it would even be possible to act as if Sufism is correct and incorrect- acting does not allow for synchronic partitions the way belief does. But it seems possible to swing erratically between two ways of acting, so this reply can be avoided at least in Nurudin's case. Second, given some plausible psychological postulates, the advantages that Nurudin gets from believing will not transfer to acting-as-if. For instance, it was his wholehearted investment in Sufism that allowed him to stay up late studying; if he was motivated by his belief, and acceptance is just acting-as-if, then we have nothing to replace that motivation. Treating acceptance as a mere pattern of action by definition does not give us any resources to explain motivation. And further, exploration is already harder to motivate than exploitation, since it is by definition not expected to be rewarding[8]. But I concede that this response depends on the posited psychological peculiarities of humans. The case of imagination, discussed in the next section, will be more apt for dismissing the possibility of reducing exploring by believing to exploring by acting.

On the other hand, let's say acceptance involves some but not all of the internal prop-

[8]I owe this point to Peter Railton.

erties of belief. Which features might then separate the two? One idea is that acceptance might be conditional, ready to be taken back when necessary; we might want to make sure to keep track of all the inferential consequences of an accepted proposition. I'll come back to this as well in the next section.

In general, the connection between belief and evidence-gathering gives us a reason to be willing to take an accuracy hit in order to explore. Agents like us are motivated to gather evidence that we judge to be promising, interesting or fruitful based on our other beliefs. These experiments form epistemic projects, and can be said to rely on a shared basis of belief in which we need to be invested in order to be motivated to gather the relevant evidence. However, even if this motivational structure were not in place, we might expect that rational agents would be constrained in their experimentation by their beliefs. Even if we place acceptance between belief and evidence-gathering in some cagses, this relationship will still persist, albeit indirectly, since beliefs will affect acceptance which will affect evidence-gathering. Thus some of our beliefs can still be evaluated based on their consequences for acquiring evidence, though these consequences will be somewhat indirect.

To give up on a *normative* connection between what we believe and what evidence we collect is to invest in an epistemology that starkly separates evidential response from inquiry. This move should be a kind of last resort, since it involves relinquishing a plausible route to explaining why experiments are justified. In the next section, I'll argue that the creation of a hypothesis is a kind of internal inquiry that depends on beliefs but carries consequences for future learning.

3.5.2 Consideration and New Hypotheses

In this section, imagine that Nurudin has an unlimited capacity to control and shape his own motivational faculty, and let's set aside the doubts in the previous section to assume that for the purposes of experimentation, exploration in belief can be easily replaced with

exploration in action without epistemic cost, and with in fact with some epistemic benefit. Nurudin will go ahead and behave as if he doesn't believe that the Sufi way is morally correct in some contexts, and does in others, but when it comes to belief, he'll suspend judgment.

This Nurudin, I'll argue, will face a dilemma when it come to constructing new hypotheses. Normally (at least for an agent who isn't logically omniscient) imagination is one route to forming new theories. In particular, we don't imagine totally random possibilities out of the blue, but we use our current understanding of how things actually are, together with beliefs about what is and isn't possible, to come up with new options. For instance, an agent might start by visualizing an atom according to her current theory which involves a series of rings, and play around with the possible ring shapes, crossing them over and changing their sizes (according to her ideas about what shapes can do (innate or otherwise)). By this process, she comes up with a few alternative ideas.

Figure 3.4 is a rough schema for how this kind of constructive mind-wandering works, and Figure 3.5 describes some of the possible constraints imposed by belief. Nurudin starts out by considering something consistent with his current theory, and relaxes various assumptions in an incremental way in order to explore the (infinite) space of possible hypotheses. The stars represent possibilities which he has fixed on as salient alternatives to the current theory. Nearness here stands for some kind of conceptual similarity - this might be constructed by the search function and hence subjective, or objective relations among the propositions.

Now, the original 'Conflicted Nurudin' had two starting points in the space, the belief that God exists and his other related theological commitments and the more politically minded, doubt-ridden cluster. Some of the epistemic benefit that Nurudin gets from having the two incoherent theories is that he can explore the space of possibilities in two very different ways that advance both projects. What if his starting points are not beliefs but acceptances or suppositions?

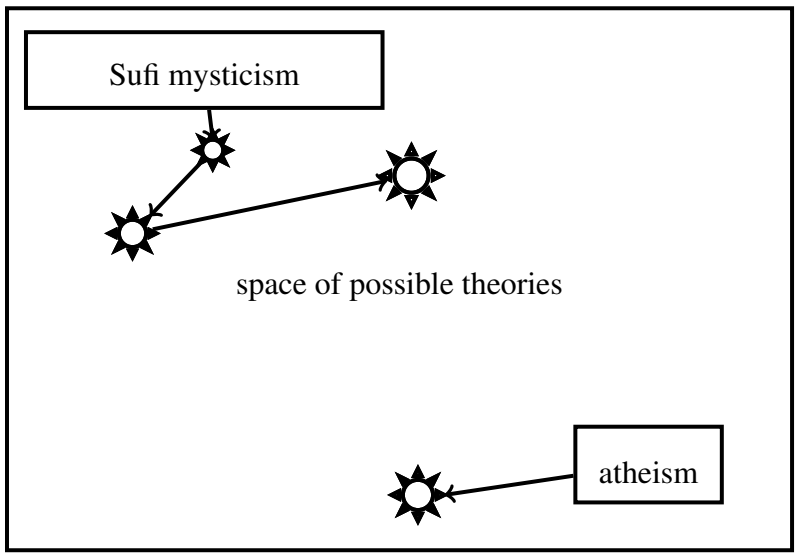

Figure 3.4: Nurudin's wandering consideration. As he explores the space of possibilities from two different starting points (rectangles), he encounters different propositions of the sort listed in Fig. 3, depicted here as stars. The process of moving around this space has some degree of randomness, but also reflects the generalization that positions nearer to starting points are easier to explore and attract more attention, ceteris paribus.

points of contact between belief and imaginative search

1. Starting points: Imagination often involves starting with our current beliefs about the world, and departing from them incrementally. So in general, we explore neighboring theories first.

2. Side-constraints: Imaginative exercises often involve coming up against previously implicit limits - for instance, in thought experiments when we observe an unanticipated reaction in ourselves that shapes the way we fill in a case.

3. Goals: Imaginative search is a combination of pure random wandering and goal-oriented activity, and the goal-directed end of the continuum depends on beliefs about what's valuable and how to achieve those ends.

4. Stopping rules: For a computationally limited agent, or one interested in computational efficiency, it will be necessary to regulate the amount of search, for instance allocating more time to search when the group seems to have reached a plateau or less when current possibilities overwhelm resources to pursue confirmation of those possibilities.

5. Costs and cost analysis: Some theories cost more resources to build than others - another consideration that favors nearby possibilities.

Figure 3.5: Sketch of possible relations between beliefs and imaginative search

This sketch is preliminary, but accords with empirical data on the kind of cognitive processing that supports creativity; for instance, Yilmaz et al.[97] detail how new ideas tend to develop as variations on existing ones, and creativity is often enhanced by pushing people to adopt a divergent starting point.

An objection might arise here. In the case of some everyday theorizing, we sometimes start from a supposition rather than a belief. Or when reading fiction, we can start our wandering from the fictional cluster of ideas and move outward, perhaps filling in other details of that imagined world by varying the fictional events ('what if Dr. Zhivago hadn't gotten on that train?'). So couldn't Nurudin start from acceptance?

Here I think we can allow that while we sometimes begin our wanderings from accepted premises, in general there is an important relationship between belief and consideration that doesn't carry over to acceptance. For one, beliefs can guide our wandering not just as starting points but as side-constraints; in this case, we don't need to attend to the constraint in order to use it. In fact in the case of imaginative resistance, philosophers have debated extensively just what some of the common side-constraints might be. Acceptance, in contrast, even on a very belief-based conception, is separated somewhat from other beliefs and inferences, and so does not intrude as much into mind-wandering. Instead, to employ supposition or acceptance in imagination, we'll need to either explicitly trigger the accepted proposition itself or already be 'within the quarantined zone' so to speak. In other words, in imaginative resistance, we have trouble turning off some of our existing beliefs about, say, morality. Acceptance, by definition, is easier to turn on and off than belief, so does not intrude as much into the exercise of imagination, including the use of imagination under discussion: building new theories.

3.6 Condition 4

I suspect that the simplifications necessary to treat belief as an iterated problem are of a kind with the simplifications necessary in the case of action - in neither case is there literal

repetition, but it's close enough for the idealization to be useful in action, so it should also be for belief.

One wrinkle is that actions like instances of pulling a lever are obviously segmented, whereas instances of believing are hard to separate from one another. How often am I in the position to say that I've believed in God seven times? But this feature is not significant for the analogy - what is required by iteration is that the same belief problem arises over and over again, so that the agent can vary her behavior or keep it consistent. Even though beliefs themselves are not properly segmented, we can categorize the evidential situation as segmented and repeatable just as in the case of action.

3.7 Objections, Alternatives

In this section, I consider objections and alternatives to my account of the trade-off for belief.

One significant challenge is that if we accept the rationality of the exploration/exploitation trade-off in action, positing an additional trade-off for belief amounts to two solutions to one problem, where each solution is on its own sufficient. That is, isn't introducing exploration twice overkill?

There's something undoubtedly correct in this suggestion - agents who introduce arbitrary oscillations, randomness or other exploration behaviors at multiple points face a difficulty in making sure these interventions are consistent. In some situations, introducing exploration at the level of action will be enough to reduce the agent's chance of getting stuck in a local maximum. And likewise for imaginative search; if we introduce randomness into the search process itself, that will solve some of the problems of a purely exploitative approach. Note that this move still involves changing the canonical framework for exploration in action, since we are interested here in the epistemic rationality of actions.

However, this will not always be the case, and there are benefits to belief exploration which do not carry over to imaginative exploration. Consider how it is that Nurudin's

beliefs allow him to explore in Figure 3.4. It's not just that he happens to explore theories that are adjacent to his beliefs; these theories are *made more accessible* to him by his beliefs. Because he believes in Sufism, through coordination of actions, imagination and other modes of thought, he's amassed resources to understand that theory and how it might be altered to create new versions. For one not familiar with Sufism in that intimate, thorough-going way, it wouldn't be clear, for instance, that there are two versions of the view, one which takes the mystical state of oneness with God to have content, and one which doesn't. Given this, in order to gain the advantage of the incoherent package by only changing actions, there would need to be a coordinated exploratory change to both external actions and imaginative ones. Changing the underlying beliefs is a natural and effective way of achieving this coordination. In other words, fully inhabiting the Sufi framework is necessary for exploring these fine-grained questions about divine experience that bear little to no relation to action. Further, even changing external actions and imagination in a coordinated way would likely be insufficient; part of how belief makes regions of possible space accesible is intrinsic, coming from the fact that believing in something involves entertaining that proposition fully, in a way that seems deeper than other forms of non-doxastic consideration.

Another objection is that my argument, and perhaps the view itself, presupposes epistemic consequentialism. Epistemic consequentialism is the controversial view that epistemic rationality reduces to a decision-theoretic problem where truth, accuracy etc. is assigned some kind of utility. I'd rather not take a stance on this issue here, so I'll note a way for an epistemic deontologist to accomodate the rationality of exploratory beliefs, and one way for an epistemic virtue theorist to accomodate it. These are not meant to exhaust the possibilities, but merely demonstrate the flexibility of the account.

Epistemic virtue theory could hold that exploratory belief is the expression of an underlying virtue or skill - something like creativity or open-mindedness. On this view, the kind of story I've sketched can be used to show that open-mindedness as a virtue is unified with

other epistemic virtues. But primarily, the function of telling this story about the trade-off serves to describe what open-mindedness looks like and how it can be distinguished from other features of epistemic rationality, namely whatever goes into exploitation. On a responsabilist virtue-theoretic picture, open-mindedness might be its own valuable characteristic, whereas on a reliabilist virtue-theoretic picture, the argument I've given in this paper shows how exploration is a reliable practice.

An epistemic deontologist is canonically not interested in justifications based purely in the results of believing in some way. They could allow for exploratory beliefs by appealing to other considerations beyond generating the right results, usually something like conforming to epistemic requirements. These requirements themselves cannot be justified by their results, otherwise we have rule-consequentialism (which is probably the most natural fit with the argument given in this paper). One non-consequentialist justification for a requirement to explore might be that trying out new beliefs is an intrinsic part of being epistemically responsible. The possibility of getting stuck in a local maximum, just like the possibility of hurting someone with a negligent bit of landscaping, would thus dictate responsible behavior even if the agent were not actually at a local minimum or her garden did not actually hurt anyone. Exploration is not a black-box reliability machine like using a crystal ball; it's a practice that's integrated into and regulated by our other ways of believing, and the account I've given here shows how we are always navigating the exploration/exploitation trade-off as we move through the process of learning.

Finally, it makes sense to ask what the upshot of this account is for epistemic rationality; does this mean that exploration is the only rational way to believe? I aim to have established a weaker thesis: exploring by believing is sometimes epistemically permissible. This follows if we assume the following:

> **Optimality Thesis**: if S believing that P will probably lead to an optimal outcome in terms of accuracy and/or truth, and S knows this, it's epistemically permissible for S to believe that P.

This thesis is controversial since it allows cases of self-fulfilling belief to establish permissability. For instance, if my belief that I will succeed in the exam is part of what makes it likely for me to succeed in the exam (by, say, increasing my confidence and thus my performance), the optimality thesis tells us that it's permissible for me to believe that I will succeed. Objections to this result are often motivated by evidentialism, roughly holding that self-fulfilling beliefs are not based on evidence in the proper way and so are epistemically impermissable (see [93]).

Exploratory beliefs share a feature with self-fulfilling belief: both cases use the (expected) results of believing in order to justify believing in the first place. But they are different in the following way: self-fulfilling beliefs make themselves rational by making the proposition under consideration true. They are only rational *once they are believed*. On the contrary, exploratory beliefs do not make any change to the truth of the proposition under consideration, nor do they make themselves rational in a causal sense. They are permissible because they lead predictably to good epistemic consequences, but in the standard way. The weirdness of self-fulfilling beliefs is this non-standard, non-ratifiable way, which is not shared with exploratory beliefs. [9] So we can amend the thesis as follows:

> **Optimality Thesis***: if S believing that P will probably lead to an optimal outcome in terms of accuracy and/or truth *in the standard way*, and S knows this, it's epistemically permissible for S to believe that P.

It's beyond the scope of this paper to reformulate the optimality thesis to reflect this difference properly - specifying the way in which self-fulfilling beliefs are non-standard or circular is a complicated project that requires a comprehensive survey of the variety of ways in which self-fulfillment works. But I take it to be sufficient in this context to point out that the difference between self-fullfilling and exploratory beliefs is precisely the feature which makes self-fulfilling belief look epistemically questionable.

[9] In other words, if we consider beliefs to be acts, exploratory beliefs need not occur in a context that violates act-state independence.

3.8 Conclusion

Our country song asked:"how am I ever gonna get to be old and wise, if I ain't ever young and crazy?". In this paper, I've argued that this same line of thought applies to belief - in the beginning of inquiry, we often should believe in order to explore rather than to exploit, but as inquiry progresses, we should drift towards maximizing evidential value. This is a feature shared between action and belief, and exploits the rational connection between belief and imagination.

An implication is that just as in the practical case where reward variability modulated the trade-off, this analysis of belief gives us room to make a parallel move. Epistemic pay-offs surely vary, and often in a predictable way. I need the right theory more urgently when I'm starting to build my machine or about to go on an expedition. At other times, such as idle inquiry, preliminary stages, or even after the plans for the machine are all in place, the stakes are lower. The framework I've put forward would allow us to say that the epistemically rational behavior depends on the pay-off - and tends toward exploitation in the high risk case and exploration in the low risk case.

In some sense, what I've said here is reminiscent of talk that motivates moving away from belief towards acceptance and other belief-like states. However, by demonstrating a symmetric trade-off in the case of action, I hope to have pushed back against this project. If the exploration/exploitation trade-off is a ubiquitous feature of goal-oriented rationality, then rather than classifying exploratory belief-like states as forming a separate category, we should expect the trade-off to occur over states of a single type. Further, by treating the phenomenon as a trade-off in the rationalization of a single state (i.e. belief), my theory has an advantage in terms of parsimony and strength . In other words, my opponent must explain how beliefs and acceptences combine in regulating behavior during exploration, and this may be a difficult task.

My view is also more flexible in describing the gradient of rational grounds as a modulation of the trade-off, since any mixture of rational grounds for a single proposition in an

acceptance-based theory can only be described by the unfortunate scheme $X\%$ acceptance, $1 - X\%$ belief. In other words, it's hard to imagine what it would mean to half-believe and half-accept something, whereas it's easy to see what it means to have a belief that results from being 50% or even 21.87% exploratory. The trade-off I have proposed is naturally graded in a way that matches the underlying normative fact that our circumstances give us reason to explore to varying degrees, shifting over time.

More generally, the choice between acceptance and belief as the states at stake here rests on what we think belief is fundamentally about. On one view, belief is the state that we use in inquiry: it guides us in performing experiments, and in dreaming up new theories. Belief is the state linked most closely to our response to evidence. If this is our picture of what belief does, then we should not choose a normative framework that starkly separates belief from experimentation and imagination.

Finally, these results shake up the classic debate in epistemology between James[51] and Clifford[27] over the reasons to believe. To simplify somewhat, Clifford's position was that belief should always be based on the evidence, and undertaken for pure epistemic reasons, whereas James held that we should sometimes believe beyond what the evidence can support for practical reasons. The exploration-exploitation trade-off for belief suggests that there is a third position available here. Sometimes, purely epistemic reasons that have nothing to do with our values, desires or interests support beliefs at odds with the current evidence.

CHAPTER IV

Memory Anomalies

ABSTRACT. In this chapter, I investigate the role of novelty in memory processing, and argue that selecting for (the right kind of) novelty in memory is a distinct rational computation. I first discuss an intuitive example of how novelty can be especially memorable. Then, I compare psychological work on novelty in large-scale theories of cognition, as well as localized in memory. While there is a wide consensus that novelty is tracked and used in some way, both in memory and in other kinds of cognition, what we do with novelty and what kind of thing counts as novelty is deeply controversial. I argue that this evidence loosely suggests a specific role for novelty in memory, and then provide a conceptual account of how such a computation would work. The chapter concludes with an argument that this novelty computation is rational, and yet goes against some of our assumptions about rational cognitive processing.

Introduction

This chapter aims to make modest progress on understanding the rational relationship between *acquiring evidence* and *building a model of the world*. Typically, we think of this relationship as flowing from left to right; we acquire evidence, and in response, alter our model of the world. However, the reverse direction is also part of the story; we acquire evidence based on our model of the world, that is, we seek out information based in part on prior expectations about where new information might be found and how to get to it. Putting these together creates a cycle where each direction might in principle constrain and effect the other - a loop, in other words. The aim of this paper is to ask, in light of this loop, does the fact that we use our models of the world to guide evidence acquisition constrain

how these models should look? An affirmative answer to this question asserts that there is a forward-looking, evidential rationale for holding a certain view of the world. Taking our memory systems to be in the business of maintaining these views, I will present a kind of cognitive operation performed by the memory system which can only be explained by appealing to such a forward-looking rationale. This phenomenon, the selective memory for anomalous events, will be an instance of the rational loop between acquiring evidence and building a model of the world: a way of storing information that makes sense in light of how it drives us to get new evidence.

I'll start with an intuitive example before surveying some relevant findings from psychology. Consider the following situation:

> **Offensive Question**: Azadeh was at a talk. Her friend, who is kind and even-tempered, asked very aggressively why the speaker employed such a terrible methodology. At the time, she thought to herself: "I can't possibly imagine what he was thinking! There's no judgment that I could make about a talk that would lead me to ask that kind of question". This incident stuck in her mind long after she had forgotten all the other questions at that particular talk. But she did not revise her judgment about her friend's character or the deeply offensive nature of the question - instead, she felt uneasy about what had happened and regarded it as an unexplained fluke. Later, Azadeh and her friend talk about what happened, and she comes to understand what had upset him, and why asking the question was consistent with his good character; he needed to publicly register his disagreement in order to demonstrate that the speaker did not speak on his behalf.

Azadeh's attitude from the time of the talk to the time of the later discussion can be characterized in the following way: she remembered what had happened, but did not incorporate the evidence furnished by the incident into her other beliefs about her friend's character. That is, in ruminating about how offensive her friend's question was, she seemed

to believe that the event occurred. However, by resisting following these thoughts out to their usual conclusion and reducing her confidence in his moral character, Azadeh seemed to not really believe. You could imagine that this resistance is active, driven by a conscious desire to avoid judging her friend negatively, but more often it might be subconscious, a path her thoughts tend not to take, the basis of which need not be luminous to her. I'll refer to memories that we treat in this way as anomalous memories. While there is a component to this example that goes beyond memory into perception, action, and attention, I take it that there is also a distinctive kind of memory involved. These memories are distinguished by their evidential attitude in the following way:

> **Anomalous memories**: A memory M of event P is treated as an anomaly when the agent singles it out for selective consolidation and retrieval based in part on the fact that P is surprising, unlikely, or odd, and when the agent is at least sometimes not disposed to draw inferences (whether conscious or unconscious) based on M because of P's oddness.

In Azadeh's case, the event was surprising since it deviated from her expectation of her friend's behavior. Because we might have more or less firmly set expectations, deviations from expectation can mean a variety of qualitatively different things. For instance, I once remember seeing a dog in a cafe that raised a single eyebrow at me. I don't believe that I antecedently thought that dogs cannot raise their eyebrows. In fact, it had never occurred to me to consider whether dogs have eyebrows. Even if I considered the question, I might have been in doubt. But the real phenomenon was just so bizarre and human-like that it stuck vividly in my memory. So at this point, I will leave 'surprising, unlikely, or odd' in the description above as deliberately vague. If we make progress on the question of the rational function of anomalous memories, it should be possible to be more specific about what kinds of events should trigger this response.

You might notice that anomalous memories have a similar structure to the kinds of scientific anomalies discussed most notably by Thomas Kuhn[55]. For instance, the sci-

entific community treated the procession of the perihelion of Mercury, a piece of data that famously did not fit into Newtonian mechanics, as an anomaly. For Kuhn, the chief function of anomalies is to accumulate until the collective weight of the evidence furnished by the anomalies provokes some scientists to jump ship from the prevailing framework. Of course, in order to accumulate, anomalies must be recorded and classified, but the focus of Kuhn's account is on what happens in response to the anomalies, not how the anomalies are generated and maintained. Kuhn would be loathe to describe anomalies in these cognitive terms, and might be even more unhappy with the flatfooted, rationalistic treatment they will receive in this paper. And so while there are significant differences in both the descriptive and normative elements of memory anomalies and scientific anomalies, I note the similarity here in order to convince the reader that understanding memory anomalies is connected in potentially fruitful ways to broader questions in epistemology and philosophy of science.

This chapter will consider the rationality of anomalous memories. That is, where does this cognitive operation fit into the behavior of an epistemically responsible agent? Might we be irrational to disregard anomalous events in memory, and if so, is this the same kind of irrationality that occurs when we believe contradictions or deny a deductive implication of our beliefs? These questions are a window into a deeper question: what role does memory have in epistemology?

I'll argue that generating and maintaining anomalous memories is a *distinctive rational function* of the long term memory system (hereafter, I'll use memory to refer to long term memory unless otherwise noted). This claim has two parts. First, this kind of cognitive behavior is epistemically rational, meaning it makes sense from the perspective of designing an agent that aims to get at truth about the world. Second, it is a cognitive operation that is particular to the memory system, meaning that the operation itself is a memory operation as opposed to a computation that just happens to take place in the memory system. And so these jointly entail that memory itself has a distinctive, active epistemic role, con-

tra the explicit views espoused by Senor[81], and the implicit consensus whereby memory is a topic receiving minimal attention in (contemporary analytic) epistemology. A second consequence is that this operation is a rational step *away from coherence*, which raises a challenge for coherence-based approaches to epistemology such as some versions of evidentialism, though I'll argue that a diachronic coherence approach can capture these cases perfectly.

In Section 4.1, I start with a brief survey of empirical work on the cognition of anomalies within the memory system as well as in online cognition. In Section 4.2, I argue for the distinctiveness of novelty within memory, and in Section 4.3, I present a series of considerations in favor of the view that maintaining and generating anomalous memories is epistemically rational. Section 4.4 concludes with a discussion of what these cases show us about the value of coherence.

4.1 Cognition of Anomalies

An anomalous memory, as I use the term, refers to a kind of memory representation (a representation of an odd event), and a fairly specific way of treating that representation (attending to it and yet keeping an evidential distance). Whether the anecdote I provided translates into a general cognitive phenomenon is hard to determine empirically, given the particularity of a case like *Offensive Question*; what strikes Azadeh as bizarre would not necessary even be noticed by another conference attendee. However, there is a broader body of literature on the ways in which novelty is attended to, avoided, and remembered.

The initial focus of this section is on contextualizing this anecdotal description within this broader area of work on the role of novelty in cognition. I first discuss novelty in large-scale models of the mind, and then in the more specific domain of memory. Novelty, as a preliminary gloss, will mean something like unexpected stimuli.

4.1.1 The Big Picture

Novelty has recently been the subject of several large-scale theories of cognition, in humans, other animals, and artificial agents. Consider the following three research programs:

1. Singh et al. encoded their reinforcement learning agents with a preference for novel stimuli, and showed that in the contexts studied, the novelty-preferring agent performs better than an agent without novelty preferences even according to the aims of the non-novelty-preferring agent. Novelty in this model means environmental stimuli which do not match the agent's expectations. The authors posit that the intrinsic desire for novelty promotes learning and thereby goal satisfaction across a variety of contexts. [23]

2. In the literature on control and motivation in humans, expectation violations are commonly taken to be per se aversive. For example, Proulx et al (2012) write: "any given inconsistency is understood as evoking a common syndrome of aversive arousal" where inconsistency is defined as any "detected expectancy violation"'. They posit a general aversive mechanism for both high and low level inconsistencies at least partly localized to the anterior cingulate cortex (ACC). [71]

3. Clark (2013) thinks of perceptual systems as functioning primarily to generate and then minimize prediction errors - i.e. mismatches between expectations and reality. These systems work in loops of feed-forward and feed-back whereby expectations are generated, fed down the hierarchy, compared with stimuli and then errors are fed backwards and used to adjust future predictions and cue attention and other behaviors. [26]

Of course, Proulx et al are offering a descriptive account about humans, whereas Singh et al are offering a normative account about rational preferences in a broad class of agents, and Clark is doing something in between. But arguably, even the more descriptive accounts have a normative output; they describe a highly general way of engaging with the world in

species that perform their cognitive tasks in at least a reasonably sufficient way given their limitations. Given that, it seems that these positions are potentially in tension.

Based on these research programs, you might be tempted to conclude that agents should (where "should" is meant in a very weak sense of "would be acting reasonably to"):

1. Search for, and value, novelty

2. Avoid, and disvalue, inconsistency

3. Generate, and then minimize, prediction error

But of course, the catch: the descriptions given of novelty, inconsistency and prediction error by these authors are incredibly similar, and at times even identical. They are all violations of expectations. And note that value differences in the stimuli cannot explain the difference (i.e. good surprises are novel, bad surprises are inconsistent) because all of these accounts posit that value is coded *in response* to surprising value-neutral phenomena. That is, they hold that novelty is per se encoded as positive - or negative. The reason these accounts assert anything interesting or controversial is because they offer an explanation for a basic way in which value can be computed.

This disagreement is about the cognitive function of novelty, as well as the rationality of that function. While it might be theoretically possible to reconcile these accounts, I'll instead treat this disagreement as motivation to look at more fine-grained functions for novelty. That is, if we've reached an impasse about the function of novelty in cognition in general, perhaps it can be resolved by assigning different functions for novelty in different cognitive modules or subsystems.

4.1.2 Novelty in Memory

As in the work surveyed in the previous section, memory for novelty is usually understood in terms of *expectation mismatch* [15] [74]. Further relevant distinctions can be

drawn to separate a broadly unexpected observation from mismatches between an observation in a context[14] or in an association[67]; for instance, a chair is not a novel stimulus in general, but it becomes novel when found in an aquarium, or associated with strong positive emotion.

When locating novelty processing in the memory system, there are several possible points in the process of memory that we might be interested in. Following a (somewhat controversial) convention in cognitive psychology and neuroscience, the processing from observing an event to later recalling it to mind can be split up as follows in chronological order: encoding, consolidation, retrieval, reconsolidation, retrieval,... with the last two stages potentially iterating many times. Encoding refers to the initial storage of information. Consolidation is the first transfer of information within the memory system, which often results in changes to the content; temporarily encoded information is moved and combined to be stored for the longer term. Retrieval involves some of the information stored in memory becoming available to online cognition. Following retrieval is a process of reconsolidation which can involve both changes to the retrieved information as well as changes to other stored information[62]. As each of these stages involve discrete computations, selecting for novelty could in principle happen at any or all of them - in this section, I will focus on encoding and (re)consolidation, since these stages have been linked more robustly to novelty.

Just as in the general cognition case, the psychological literature on novelty in memory reveals an immediate problem: mismatch is sometimes remembered more than 'match'[14], sometimes less [94], and sometimes there seems to be no effect at all[66]. However, this plausibly reflects differences in the kind of novelty at stake; for instance, Rouhani et al.[78] found enhanced memory for any reward prediction error regardless of valence. Reward prediction error refers to a mismatch in how much value obtains in a particular state as opposed to other features of that state. In the classification task of identifying novelty which should be remembered and novelty which can be put aside, features such as high contex-

tual stakes might suggest that this odd event should be given extra attention, whereas other features such as the novel observation being only novel in a particularly narrow context, might instead point to the odd event as an outlier or measurement error that should be de-emphasized or even forgotten. We seem to be far from a comprehensive answer to this question. I will mainly focus on the cases of *enhanced* memory for novelty in order to at least sketch the known effects within that category.

At the neurobiological level, researchers have uncovered several signatures of novelty-associated computation. The increased salience of novelty at encoding seems to be modulated by the neurotransmitter dopamine. For instance, Rangel-Gomez et al. found that treating humans with the dopamine agonist apomorphine increased the strength of memories for words written in novel fonts relative to those written in regular fonts. This task involved testing shortly following the event and correlated with an ERP (event-related potential) previously associated with novelty at the encoding stage[75]. On the other hand, while dopamine is thought to be enhanced in schizophrenia, Mayer et al. actually found reduced novelty salience at encoding in schizophrenic patients[61]. Their task involved a novel upside-down letter 'A' compared to a non-novel properly oriented 'A'. At consolidation, work using fMRI has linked novelty at encoding to increased activity in the medial temporal lobe, relative to increased cortical activity for non-novel stimuli[92]. Here, a representative experiment[92] involved a rubber duck in ordinary (bathroom) and novel (bakery) contexts. As you can see, these three studies used very different stimuli under the heading of novelty. The sense in which an upside-down 'A' is odd is quite different from how a rubber duck in a bakery is odd. For one, the former is novel in general and the latter is only novel in the context. And even the relatively small difference between a word in an unusual font and an upside-down 'A' might explain the divergent findings in the first two dopamine studies.

A key question for novelty processing in memory is how much of what ends up selectively encoded as novel in memory is inherited from other cognitive systems as novel? It

is safe to assume that at least some observations are passed down in this way, given other work on close links between working memory and encoding into long term memory. An object tagged as novel in perception may be recorded as novel in memory without substantial intervening computation. This suggests three potential categories where novelty arises in either or both of on-line cognition or memory: (a) observations tagged as novel by on-line cognition and consequently treated as novel in memory, (b) observations tagged as novel by on-line cognition but treated neutrally in memory, and (c) observations treated as neutral on-line but tagged as novel in memory. Category (b) can be given an efficiency rationale. Murty et al. [66] note a computational trade-off between recording every novel event that might be salient to reward, or taxing the memory system less and potentially missing salient cues. In their study, novel stimuli were selectively remembered in only some environments: namely, those with salient rewards.

Categories (b) and (c) both involve a computation by the memory system, in particular, (c) requires an active computation to tag new stimuli as novel. Whether these actually characterize typical computations in the mammalian memory system has not yet been firmly established. However, the resources required to do such a computation are already invoked by most theories of consolidation, which appeal to the notion of a schema or similar concepts. A schema is something like a script or pattern that is used to encode what has happened and predict what will happen - for instance, a narrative structure that says folk stories should have a beginning, crisis point, and then happy ending. A Baba Yaga story, for example, where the witch Baba Yaga menaces a child but is then dispelled by a clever trick is schema-consistent. There is a strong consensus that memory processing involves selection for schema-consistent information, and that this process is ongoing from encoding to reconsolidation (see McKenzie and Eichenbaum[62] for an example of one such theory and overview of others). Presumably, if our memory systems have the computational resources to selectively enhance schema-consistent events and stimuli, these same process might be modulated to highlight certain schema-inconsistent events and stimuli. Of

course, we can only speculate at this juncture about what such a computation might look like, but the evidence for some kind of novelty-sensitivity in memory that I have surveyed gives us an additional reason to search for memory computations that select for novelty.

In summary, empirical evidence indicates several possible roles for novelty. While much of the research emphasizes the importance of novelty, large-scale theories disagree dramatically about how novelty is treated. Research in memory echoes a similar clash, though features such as risk, reward and scope of expectation seem to modulate the response to novelty and possibly provide clues as to how to reconcile the cases where novelty is more memorable, neutral or less memorable. It is plausible, though far from established, that memory encoding, consolidation and reconsolidation involve a computation selecting for novelty, and actively making some novel events especially easy to retrieve.

4.2 A Distinctive Function

In this section, I argue that some possible novelty computations should be thought of as distinctively memorial operations. That is, computing these functions would consist in performing a memory operation. Where the previous section motivated a search for this kind of computation as a biological fact, this section will sketch a theoretical model of the relevant computations. Here, I distinguish memory operations from on-line operations as follows: memory operations involve time-scales of days or longer, they do not typically involve conscious attention, and in general concern information storage. In the experimental setting, memory operations can also be distinguished neurophysiologically, though I will not take this approach here.

Generating and maintaining anomalous memories is a distinctively memorial operation given that anomalies in memory have **distinct content** and **distinct roles** from anomalies in attention and on-line cognition. This section makes a conceptual point about the intuitive category of memory operations, not an empirical point about how we actually compute these operations; the aim of doing so is just to set up the phenomenon which will be given

a rational explanation in Section 4.3.

Building on the example in *Offensive Question*, we should expect anomalous memories to sometimes have distinct *contents* from anomalies in on-line cognition. There are two ways in which such a divergence might occur. Some perceptual anomalies do not make it to long-term memory as anomalies. For example, a weird splotch that draws visual attention turns out to be a speck of dust on my glasses, and I do not encode it vividly in memory. Conversely, some memory anomalies may not be tagged on-line, but instead emerge later. For example, a normal-seeming sighting of a friend becomes anomalous when I later hear that he was supposed to be in Bosnia at the time. These two kinds of mismatch in content between memory and on-line cognition correspond to the categories (b) and (c) from the previous section.

Considering how anomalies can be used, we can also see a distinction in *function* between memory and attention. In *Offensive Question*. Azadeh holds on to this peculiar instance, and as a result, examines her future interactions with her friend more closely. That is, not unlike in the case of our Kuhnian scientist, she is looking for further anomalies to record and combine together. Unlike in the simplest Kuhnian model where the anomalies just push the community one notch further toward a paradigm shift, in a realistic case, anomalies need to be remembered as well as processed through the probability distribution to avoid repeats and to be properly chunked. Realistically, anomalies that perturb a scientific theory will not be an uniform series of events, but belong to clusters and structures that the agent would benefit from attempting to track. For instance, a series of phenomena that link time to perspective should be clumped together in order to see how they challenge the Newtonian paradigm. This process of accumulation followed by sorting would take place over days, months or even years. Even after the paradigm shift, some of the anomalous events should be processed again; they need to be assessed on whatever new theory has arisen. This recursive and temporally-extended memory process applies to the non-scientific case as well. For instance, once Azadeh comes to understand her friend, she

will look back on his behavior and re-analyze it under the new theory, seeing his facial expressions as signifying different emotions and so on.

Thus we can conclude that anomalies could be tracked in memory in order to sort, cluster and recombine them to discern the structure of the underlying phenomena and guide behavior. Likewise, anomalies are used to prime and modulate attention over long timescales. The memory of an unusual event, as in the example, can lead us to seek out new information, sometimes by changing behavior and other times by directing attention. This indicates a role prior to behavior and attention, keeping the anomaly in the background so as to be able to use it in on-line computations. Of course, long-term guidance of attention requires more than just memory; my point is that it *critically* involves memory. In other words, anomaly in the memory system has a distinct role to play.

As an aside, anomalies may also have distinctive content and roles in the sub-categories of semantic and episodic memory. Consider the memory for anomalous states (for example, seeing a new configuration of lines) as opposed to anomalous acts (for example, attempting a new style of dancing). Observations in either category can in principle be recorded episodically or semantically. This is especially true of anomalous states, since they can be updated based on information the agent knows without information about where that knowledge comes from. There may be advantages to each kind of encoding. However, memory for anomalous acts are different, in that they are most naturally recorded episodically since they need to be updated based on information about what the agent herself has done.

4.3 A Rational Function

The preceding portions of this chapter described a behavior of selectively remembering odd events, using that memory to guide future actions such as searching for more information, and so forth. I'll now argue that this behavior, given the right conditions, underlies epistemically rational changes in belief. Rather than commit to a theory of rationality, I'll

provide reasons to think anomalous memories are rational on several different understandings of rationality. Because anomalous memories as described invoke computation and processing as opposed to just transitions between beliefs, and most theories of epistemic rationality concern only beliefs, this section will attempt to sidestep this issue by talking about the changes in belief that go along with anomalous memories, on the understanding that there is more to the computation than just a change in belief. In other words, since computing memory anomalies makes a change to the agent's model of the world, I will treat this computation as if it is a behavior consisting of belief formation and revision, since at the least it has implications for belief formation and revision, as is clear from the definition of anomalous memories. I'll now turn to a few sketches of what it means to be epistemically rational.

4.3.1 Closer to Truth

Some epistemologists are epistemic consequentialists in the following sense: they hold that change to one's beliefs is rational if and only if it brings the agent predictably closer to truth. 'Predictably' here can mean several, critically different things; the most obvious of these are 'according to the agent's perspective', or 'according to reliable processes'. The former invokes how things look given the agent's evidential position, whereas the latter is sensitive to facts beyond the agent's view. [1] I will consider both the agent's perspective version of consequentialism and the reliablist version in turn.

First, the computation of anomalous memories may seem likely to result in truth from the agent's perspective in cases where she believes that holding on to the memory of this event over time will direct her attention and result in success at eventually getting to the truth about the event. For instance, Azadeh might think to herself: 'I can't figure out what's

[1] 'Ex-ante' is sometimes added to the agent's perspective to disambiguate two ways of evaluating a change to one's beliefs: ex-ante, i.e. before adopting the change, or ex-post, conditional on the change having been made. Ex-ante and ex-post evaluations can differ in cases where holding a belief either justifies itself, or undermines itself. For instance, in a case introduced by Jennifer Carr[21], my confidence that I can achieve a handstand increases my actual capacity to do a handstand. In this case, a belief that I will definitely succeed at the handstand is ex-post justified, but may not be ex-ante justified.

going on with my friend right now, but I'll resist the urge to revise my beliefs and come back to this later, when I'm more likely to figure out what happened'. She must also judge that believing the event happened while not following through completely on the inferential consequences is her best shot at getting things right in the meantime. This may be because she doubts that any inferential consequences she draws now will be correct, and they might even interfere with figuring out the puzzle. This version of consequentialism is quite easy to satisfy, and we need not establish that it will hold for all anomalous memories but only a subset in order to show that the phenomenon is generally rational in some contexts.

Likewise for the reliablist version. Reliablist are happy for relying on fortune-tellers and magical oracles to sometimes be rational, so it should not be too hard to establish our conclusion. Further, the practice of anomalies in science lends support to the idea that preserving anomalies is a broadly useful process. The fact that we likely carry out this process as a matter of empirical fact, as I've discussed, is another line of evidence that it likely has some use to us. And the most obvious kind of use of information-management is to build helpful representations for achieving our goals. Since accuracy is systematically (though not always) helpful, this provides a reason to suspect that a widespread cognitive process without any obvious defect may be furthering the accuracy of our representations. The reliablist might also appeal to work on the temporal dynamics of convergence in group epistemology. Zollman[99] for instance presents a model on which excessive connectivity (i.e. information sharing) between individuals leads to overall worse outcomes as compared to a mild amount of modularization (i.e. partial information segregation). Essentially, this is because these connected systems converge faster, and so may more likely converge on the wrong answer and stay there, whereas systems that wait longer to converge are less likely to converge on the incorrect answer because they have accumulated more total evidence. Arguably, the same model can be applied to anomalous memories; here the limits on drawing inferences based on the memory act as barriers to information transfer, and so the agent who maintains anomalies takes longer (and waits for more evidence) before

making up her mind about the consequences of what happened. In noisy cases, where the possibility of false convergence is real, this computation should reliably increase truth over time relative to a strategy that involves faster reconciliation.

4.3.2 Apt Response to Evidence

To move on to a more substantial claim, I'll argue that anomalous memories can be apt responses to evidence. This notion of epistemic rationality, unlike consequentialism, is backwards-looking in the sense that it demands that beliefs conform to what the agent has observed rather than what will help her out in the future. By apt, I mean something like fitting with the current evidence. The central idea is that an apt response is still apt even if it is useless or even harmful; aptness is purely backward-looking.

Anomalous memories initially seem clearly inapt. Azadeh, by holding back from inferring something about her friend's character, is resisting what looks like the apt response to her evidence. But this assumes that aptness is *synchronic*. The synchronic assumption is necessary because we're considering just the aptness of her initial response, as opposed to the whole trajectory over time. If we instead analyze Azadeh's trajectory, the series of transitions is less obviously inapt, because she is in fact conforming her beliefs to the evidence eventually. She figures out the correct explanation of her friend's action not by luck or in spite of her original stance, but because of it, and there's an evidential story to be told about how. That is, she comes to see what her evidence supports, and responds by believing in the way her evidence demands.

This response can only succeed if diachronic evaluations need not be forward-looking. To see this, consider the difference between a plant growing *naturally* and a plant growing *successfully*. To say it grows naturally is to say that the stages of the process unfold according to biological regularities; it opens its flowers in response to light waves as a function of hormonal response, for instance. This is a diachronic evaluation because rather than talking about snapshots in time and transitions between them, it adverts to temporally-

extended features of processes. After all, opening flowers in response to light can only be assessed over the course of a whole day. On the other hand, this is not a forward-looking evaluation, because it does not determine whether the process is natural based on where it will end up. By contrast, successful growing, or growing that results in edible fruit, is a forward-looking evaluation. Rather than the shape of the processes taking place, successful fruit-bearing is determined by the connection between these processes and the likelihood of a future end: producing fruit. Insofar as different processes produce the same outcome (perhaps across various conditions, or enough times), they are evaluated to be equally successful - whereas a similar process failing or succeeding to meet the outcome will have a different evaluation. Further, the distinction still holds if we understand forward-looking evaluations as being about likelihoods of good outcomes rather than actual good outcomes. Here, due to a defective environment, one perfectly natural process may predictably lead to no fruit, whereas the same process in a different environment would likely lead to fruit. So diachronic does not imply forward-looking.

I'd like to suggest that on a diachronic understanding, anomalous memories are apt responses to evidence. Like natural growth, they are processes that unfold over time as a response to the agent's evidential situation and bring her into conformity with the evidence as they unfold. I have not argued that synchronic understandings of aptness are wrong, but they are not the only coherent notions of aptness.

4.3.3 Coherence

Finally, another view of epistemic rationality centers around coherence. Coherence gets invoked in both the cognitive science and philosophy literatures, often in somewhat distinct senses. But in both cases, coherence is often thought to be crucial for rationality. For instance, the psychologist Allison Gopnik writes:

> [C]ognitive maps are coherent. Rather than just having particular representations of particular spatial relations, cognitive maps allow an animal to

> represent many different possible spatial relations in a generative way. An animal that knows the spatial layout of a maze can use that information to make new inferences about objects in the maze. [42]

Coherence, for Gopnik, involves being disposed to use what one knows to extend one's knowledge to new cases. Already, the case of anomalous memories seems problematic; the evidential distance in these cases involves hesitance to extend one's knowledge. Philosophers often appeal to the following distinction to understand coherence:

- Consistency: two representations of the world are consistent if they could *possibly* both be true at the same time
- Coherence: two representations of the world are coherent if they could *plausibly* both be true at the same time

To make a representation coherent, however, will often involve the kind of response that Gopnik is describing; when I notice a sense in which I don't believe the plausible consequences of some of my other beliefs, I have a reason to fix something and often (but not always) this fixing will involve generating new beliefs. One wrinkle is that coherence is not necessarily a purely internal matter; we also want our representations to *remain* coherent as new information is fed into the system.

As in the previous section, I will not attempt to argue that computing anomalous memories furthers synchronic coherence. In fact, under both Gopnik's generative notion and the more austere criterion of coherence above, anomalous memories start off as markedly incoherent compared to other available responses such as suspending judgment entirely or following through on all the consequences of occurrence of the anomalous event. Azadeh believes that her friend has a certain character, and also that he asked this question. She also believes that no one with that character would ever ask that question. And so her beliefs are incoherent, and even inconsistent. Of course, synchronic coherence theories can allow for some rational incoherence. But synchronic incoherence is only synchronically justified

by a pure coherence theory if the agent has no other, more coherent option available. For instance, in the case of a paradox, we have no coherent option at all, and in other cases, we may have one less incoherent option that is still incoherent. Azadeh's case fits neither paradigm, because the options of suspending judgment about her friend's character, the question itself, or both are available to her and more coherent. I'll argue in spite of this that anomalous memories do increase diachronic coherence, sometimes uniquely.

Diachronic coherence can be thought of in three ways: as eventual coherence, aggregate coherence, and true diachronic coherence. Eventual coherence is about where you end up, whereas aggregate coherence weights coherence at some or all times and adds it together. True diachronic coherence does not weight any one time, but describes the temporally extended process. In fact, we've already seen a version of true diachronic coherence; diachronic aptness is a way of diachronically achieving coherence between one's beliefs and one's evidence. Aggregate coherence is interesting but since there are uncountable ways of aggregating, I'll focus on eventual coherence. Eventual coherence represents a middle ground between the previous two conceptions of rationality: it is outcome-focused, like consequentialism, but appeals to notions of fit, like aptness, since coherence is a kind of fit with the world.

Anomalous memories are rational, under rationality-as-eventual-coherence. Compare the possible end-points in Azadeh's situation:

1. The original case: she comes to see her friend's behavior for what it was, a response to a problematic dynamic in the Q&A

2. Suspension: she decides to suspend judgment about her friend's behavior initially, and never finds out more since the subject is out of her mind[2].

3. Commitment: she takes the evidence of the *Offensive Question* as decisive, judges

[2]Suspension here should be distinguished from lack of belief; since she is already considering the proposition, I will leave aside the option of never forming any attitude at all to her friend's behavior, which most naturally occurs in cases where we do not even consider some proposition.

her friend's character to be worse for it, and never finds out more since she spends less time with him.

If these were the three most plausible outcomes, it would follow that Azadeh's actual response does the best at eventual coherence. This is because both of the other two cases leave Azadeh open to being confronted with further evidence if she encounters her friend, and so being incoherent, or of encountering other facts that depend on the truth of the situation, which would be incoherent with her beliefs. Alternately, her beliefs would not end up being coherent if she actually does have some information that conflicts with her judgments of her friends character; that is, it might be possible to determine the problematic dynamic of the Q&A from the evidence she already has, if she were more discerning or motivated to uncover the truth. In this case, she is already incoherent, but in a way that is inaccessible to her.

This all may seem rather cheap. After all, many kinds of peculiar, patently irrational processes might end up with a coherent outcome - why should that tell us anything about their rationality? The key here is not that Azadeh ends up in a good place, but that she avoids ending up in a bad place. Her evidential situation at the beginning of the story is not ideal nor is her ability to discern alternative possibilities. Likewise, Kuhn's scientist before Newton is in a difficult position; she is starting to notice things that don't fit with her current theory, but is unable to come up with an alternative. In these cases, the familiar, synchronically coherent response leads to eventual *incoherence*. So anomalous memories may be rational because they are one of the only plausible solutions to these difficult problems. This is a preliminary indication that anomalous memories are not just rational, but distinctively rational; they represent a way of being rational (stepping away from coherence now to achieve coherence later) that differs structurally from many paradigmatic rational changes in belief.

In conclusion, this section established that under some but not all views of rationality, anomalous memories are epistemically rational. The views that sanction anomalous

Figure 4.1: An exhibit from the Museum of the Great Patriotic War in Moscow, depicting the siege of Leningrad.

memories are both consequentialist and deontological, synchronic as well as diachronic.

4.4 Upshots, and a Metaphor

In Figure 4.1, you can see a characteristic exhibit from the Central Museum of the Great Patriotic War in Moscow . As you might infer from the name, this museum commemorates World War II; housed behind and underneath, a giant obelisk, it presents artifacts and depictions of battles and diplomatic relations during the war. These exhibits promote a single, perfectly coherent vision of what happened, why it happened, and who was to blame. In diorama after diorama, peasants with earnest expressions and beautiful faces fight masked German soldiers arranged in exacting military formations and armed with weapons that cast dark shadows. In other work, I use the metaphor of a museum to describe the distinctive role of the memory system in generating knowledge - here, I'd like to point to the example of this war museum to show the dangers of a memory system that does not generate and maintain anomalies *even if it gets things right*. If I tell myself the story of my past in such a

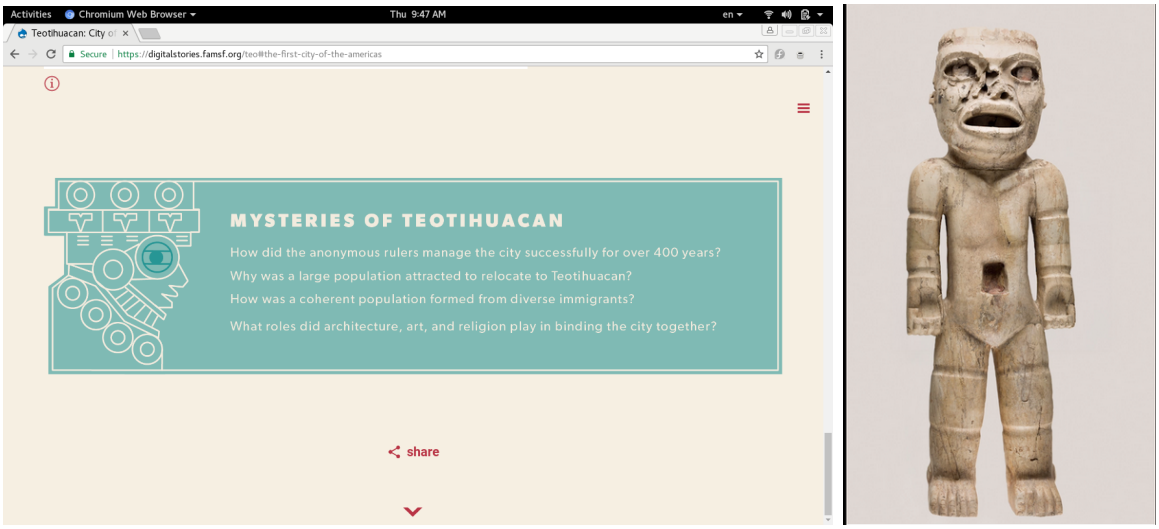

Figure 4.2: An online exhibit highlights the puzzles surrounding the fall of the city of Teotihuacan, and draws attention to a reconstructed artifact (right) which was apparently broken up into many pieces and deliberately separated across many disparate locations

perfectly coherent narrative, I am handicapped in the same way as visitors who only learn the version of history from the Patriotic War Museum. That is, if the story is in any way inaccurate, it is now very difficult for the visitors to see the flaws, and to adapt the story once the flaws have been noted. Even if the story is accurate, the viewer is not able to argue for its accuracy or discern possible weaknesses that need to be defended. This museum is not only a disservice to the facts currently at our disposal, but it has made itself nearly useless for the future when further facts might come to light.

Conversely, consider the value of a museum that presents, and highlights, artifacts and information that don't comfortably fit into the prevailing explanation, or even any explanation. For instance, the exhibit on the fall of Teotihuacan in Figure 4.2 highlights a statue that was for some unknown reason broken up into many pieces and distributed all across the city. This kind of museum invites visitors to entertain doubts about the prevailing explanation, and can serve as a hint that allows us to peel away the mistaken narrative and see a way forward to a new possibility. This mirrors both the justification I've given for anomalous memories and the Kuhnian theory of scientific revolution. As I've argued here,

memory plays a significant and irreducible role in this story.

Epistemology and scientific theories of rational cognition often focus on how we combine premises to draw inferences, bring together pieces of evidence to support conclusions, and various other instances of information synthesis. While the projects grouped under this heading differ in a myriad of significant ways, they are all ways of engaging in a broader project of studying how we combine information into a coherent whole. The contrast case is the study of how we separate out evidence, differentiate between theories, and understand experiences as anomalies, outliers or surprises. Understanding memory anomalies is a small part of the latter project, the contours of which I've begun to explore here.

CHAPTER V

Belief, and its Fraying Edges

ABSTRACT. This chapter considers memories of events and facts that in some sense look like beliefs about those events and facts, but on the other hand deviate from the content and function typically associated with belief. I use these belief-like memories to raise a challenge for the idea that belief is a natural category of mental state with well-defined boundaries. An alternative view, which has its roots in medieval theories of cognition, treats the category of belief as a small and fuzzy region in a larger, more complex space. This space is characterized by, on the one axis, degrees and types of conceptualization, and on the other, degrees and types of cognitive commitment. Returning to this classical picture of mental representations that we take as true gives us a better explanation of the wide range of states found in memory.

5.1 Introduction

Do you believe everything you remember? "I remember Tehran is the capital of Iran" seems to imply that I believe Tehran is the capital of Iran, but on the other hand, my distant memory that links the Constitutional Revolution to something about the 19th century fatwah against tobacco seems too vague to qualify as a belief. So what is the right way to characterize memories which are not beliefs, or not exactly beliefs? Can these belief-like states be held rationally, and used to guide rational action? These questions concern the intersection between two areas of philosophical inquiry: the taxonomy of cognitive states, and the rational evaluability of our mental lives.

In this chapter, I first discuss a series of cases of memories that are hard to classify in

relation to belief, starting with the case of anomalous memory from the previous chapter. I then discuss recent work by Andy Egan, Agustin Rayo and Adam Elga, and Robert Audi on belief-like states that are not clearly belief, using this discussion to make the case that these intermediate states raise some interesting questions for how to think of the attitudes we use to represent the world around us. These recent projects are interesting, but do not fully answer these questions. To get closer to an answer, I turn to a medieval theory of mental representations that does not center on belief. This account is drawn from the classical Islamic philosopher Abu Nasr Al-Farabi, who developed a distinction between conceptualization and assent. This distinction allows us to understand the relationship between all kinds of memory states and what we now think of as paradigmatic belief. As opposed to contemporary views which define belief-like states in terms of their deviation from belief, Al-Farabi's model systematizes a rich space of representational states of which paradigmatic belief is just a small region. In moving away from the belief-centric model, belief-like states can be understood in their own right as rational or irrational, as opposed to merely defective beliefs. Thus, this chapter makes a preliminary case for a return to an older, more pluralistic theory of what it means to take something to be true.

5.2 Belief-like Memory

The aim of this section is to bring to life cases of memories which are not exactly beliefs - but not obviously a wholly different state either. These cases are intended to be intuitively recognizable, and to represent several general types. However, instead of pure phenomenology, I intend to appeal to a common sense reconstruction of what is going on behind the scenes when we experience the (hopefully familiar) phenomenology of remembering.

Rather than rely on a theory of belief and argue that these cases don't fit such a theory, I'll rely on the reader's intuitive sense of what a belief is and how it behaves. For instance, consider the ordinary way I believe that it rained yesterday: I have a representation of the

rain falling on my glasses, I would say that it rained if asked, the rain fits into my narrative of yesterday's events, I would not be surprised to discover my sneakers are still wet this morning, and so on.

This commonsense function of belief can be split into three components:

1. Beliefs are formed, sustained, and revised in response to evidence

2. Beliefs guide action, including assertion

3. Beliefs guide thought, including inference and attention

By guiding action and thought, I mean that belief is involved in guidance, along with other states such as desire. I will assume that in at least in some cases, I don't infer my belief that some event occurred from the fact that I remember it (i.e. I don't compute an additional belief representation on top of my memory representation), but rather my believing that it rained *consists* in my memory that it rained[1]. For simplicity, in the chapter I will not delve into the issue of occurrent versus dispositional belief. I'll treat belief as the mental state which performs whatever we take to be the commonsense function, a state that you can be in even when not explicitly considering the content of the belief.

Consider the following case:

> **Offensive Question**: At a talk, a friend of mine who is kind and even-tempered asked very aggressively why the speaker employed such a terrible methodology. At the time, I thought to myself "I can't possibly imagine what he was thinking! There's no judgment that I could make about a talk that would lead me to ask that kind of question!"'. This incident stuck in my mind long after I had forgotten all the other questions at that particular talk. But I did not revise my judgment about my friend's character - instead, I felt uneasy about what had happened and regarded it as an unexplained fluke.

[1]This is of course not entirely uncontroversial - Sellars[80], for instance, holds that all perception involves inference, and the parallel position in the case of memory would be incompatible with the gloss I have just provided. However, for my purposes, it is sufficient that memory representations encode a picture of the world, and so the question can arise: are these always beliefs, and if not, what are they?

Later, my friend and I talked about what happened, and I came to understand what had upset him, and that he needed to publicly register his disagreement in order to demonstrate that the speaker did not speak on his behalf. But my attitude from the time of the talk to the time of our later discussion can be characterized in the following way: I remembered what had happened, but did not incorporate the evidence furnished by the incident into my other beliefs about my friend's character. That is, in ruminating about how offensive my friend's question was, I seemed to believe that the event occurred. However, by resisting following this thought out to its logical conclusion and reducing my confidence in his moral character, I seemed to not really believe. I'll refer to memories that we treat in this way as anomalous memories; these memories are distinguished by their evidential attitude in the following way:

> **Anomalous memories**: A memory M of event P is treated as an anomaly when the agent singles it out for selective consolidation and attention based in part on the fact that P is surprising, unlikely or odd, and when the agent is at least sometimes not disposed to draw inferences based on M because of P's oddness.

This evidential attitude of partial distancing may be rational in some cases. In *Offensive Question*, keeping the odd experience in the back of my mind allowed me to understand what happened later. These cases of anomalous memory parallel more familiar ones in the philosophy of science notably described by Kuhn as anomalies, such as the scientific focus on, and distance from, the phenomenon of the procession of the perihelion of Mercury in Newtonian physics. Our token Newtonian physicist was preoccupied by the procession of the perihelion, kept it in mind over time, but did not follow through on the inferences that would normally be licensed by a piece of evidence that is inconsistent with our best theory. That is, the scientist neither discounted the anomaly nor reduced her confidence in Newtonian physics.

To move beyond anomalous memories, here is a limited series of other types of memo-

ries that fit uneasily into the category of belief:

- **Distant memories**: Memories that are only triggered by a very specific, and not particularly informative cue. (e.g. Proust's childhood memory of being lost and locating himself finallly by a glimpse of the church steeple is triggered by the taste of a madeleine, and moving through a series of other recollections first. Were he to be asked, in another context, if had ever used the church steeple for navigation, he might have said 'no'.)

- **Fragile memories**: Memories that, once recalled, would be rejected as false. (e.g. I have a memory with the content 'walking without slippers gives you the flu'. However, once I call to mind this fact, it's immediately obvious that I believe this to be false, since I consider it a folk superstition.)

- **Recalcitrant memories**: Memories that are both accurate in their first-order content, and known to be confabulated (e.g. I seem to remember my brother's bris, but I know that this memory was actually formed after hearing my father describing the event in great detail, and contains many details that I would not have actually noticed at that age.)

- **Thin memories**: Memories which are extremely information-poor, requiring the details to be filled in with effort and background information (e.g. I have a sketchy memory of a trip I took as a child, but the only way to recognize the event and string the fragments into a narrative is to think very hard about where I could have been, who would have been there, and so on.)

- **De-contextualized memories**: Memories which seem to consist solely in a feeling or sensation, as opposed to an event or proposition (e.g. I have a memory of happiness but cannot remember why or what about, or a memory of a geometrical figure without any idea of whether I saw it or thought it up.)

This taxonomy is incomplete, but hopefully gives the reader a sense of the range of memory phenomena beyond the paradigmatic cases. Note also that switching from full belief to partial belief or credence does not alleviate the tension here; like full belief, credence pictures typically operate with a restricted sense of what a 0.6 credence is, what kind of evidence forms it, and what kinds of behaviors it licenses. This means that there is a functional profile in common between all credences that I hold to the same degree, whereas the phenomena I just noted could describe five different ways of treating the same proposition. Thus, these differences cannot be explained by appealing to degrees of belief.

5.3 The Problem, and Some Partial Solutions

Memories are not the only cognitive states that fall in between belief and something else. Imagination, consideration, guessing, and implicit attitudes have all been sometimes identified as belief-like but not quite belief. While I don't have space here to discuss these states at length, I will discuss the difference between the memory cases above and the case of monothematic delusions, using the example of the Capgras delusion, which has been treated at some length by philosophers.

In the Capgras delusion, a patient sees a familiar person but comes to suspect that they are an imposter, the true familiar person having been replaced through some conspiracy. A popular theory about this delusion holds that it is caused by the ability to identify faces, coupled with the inability to associate faces with affective responses. When the familiar face does not lead to the usual affective response, perhaps including the cuing of episodic details, the patient confabulates a story about the changed identity of the familiar person to explain the mismatch. While the Capgras patient in some sense believes that the person has been replaced, this view is not formed rationally, and does not lead them to take consistent rational action. Instead, they typically switch between acting under the delusion, and as if it were false, and rarely entertain the obvious thought: what if I'm being paranoid? In only a small percentage of cases does the patient resort to taking actions to unmask the 'in-

truder', and they tend to report the delusion only sporadically[83]. Some researchers have attempted to rationalize these responses as a result of a defective percept or a shift in the patient's prior probabilities. However, given that the majority of Capgras cases occur in the setting of a chronic psychiatric disorder (typically schizophrenia)[29][36], the default theory for at least this subset of the cases can reasonably appeal to defects in thought patterns. V.S. Ramachandran notably provided an explanation that both emphasizes the bizarre nature of this delusion and shows how it is somewhat continuous with normal cognition[73].

I discuss the Capgras delusion here to raise an additional explanatory challenge. Both the Capgras delusion and anomalous memories are states that resemble belief but do not share all of its features; the Capgras delusion is formed through a peculiar process and is inconsistently held as well as resistent to counter-evidence, where anomalous memories are resistant to incorporation in many inferences, and lead the subject to dwell on them beyond the ordinary. However, the Capgras delusion is manifestly irrational, whereas anomalous memories are at least plausibly not irrational. If the parallel with anomalies in scientific practice is correct, this memory phenomenon may even be a core part of rational change of view.

The possibility of belief-like states like anomalous memories and the Capgras delusion raises a series of challenges for traditional accounts of belief. In this section, I'll introduce these challenges through discussing several projects that focus on different belief-like states. These projects are Agustin Rayo and Adam Elga's fragmentation view, Andy Egan's theory of bimagination, and Robert Audi's account of dispositions to believe.

On the fragmentation view, the ordinary category of belief is understood as having an asterisk. I believe that it will rain on Sunday *relative to a certain purpose or query* - for instance, for the purposes of planning my week. Elga and Rayo introduce this notion to capture cases like distant belief, and especially those with an asymmetry in cues. In one of their examples, I have a hard time identifying an English word that ends with 'mt', while it's easy to identify what the last two letters in 'dreamt' are. This asymmetry shows that it

would be missing something to flatly assert that I believe that 'mt' are the last two letters in 'dreamt'. Rather, for one purpose (or to answer one question) I believe it, and for another, I don't.

This kind of asymmetry motivates Elga and Rayo's picture because it is both ubiquitous and puzzling on our ordinary understanding of belief. It is ubiquitous because most things we believe are not accessible relative to all purposes. The asymmetry is puzzling because it is not explained by our familiar accounts of the nature of belief, nor by folk psychology; on most views, as in the intuitive gloss provided above, belief is an attitude towards a proposition that the agent either has or fails to have. Of course, there is nothing intrinsically suspicious about cases that are similar to but not quite belief. But I take it that Rayo and Elga want the fragmentation cases to be explained *somehow*, and so folding them into the category of belief is especially parsimonious.

To relate fragmentation to the memory phenomena I've been discussing, Elga and Rayo's cases are essentially of the same kind as distant memories - accessing the content is dependent on a cue. The other memory types I described cannot be assimilated under this view. Fragile memories, for instance, are not believed relative to any cue, since once cued, they would be rejected. In summary, Elga and Rayo's account responds to this problem of belief-like states by enlarging the category of belief. The result is that one of the memory cases now counts as belief, but the others still remain in the belief-like category.

Andy Egan argues for the existence of 'bimagination', states between belief and imagination, to deal with the Capgras delusion. His argument precedes in part by appealing to fragmentation. If the category of belief is already so heterogenous as fragmentation demonstrates, why not posit intermediate states between belief and imagination? That is, fragmentation shows that the dispositions involved in believing can be *active* in a limited setting, relative to a purpose. And Egan will exploit this feature to argue that the notion of being belief-like but not belief is natural. He describes the alternatives as follows:

> "On the restrictive view, there are only a few roles available - there are a small

> number of representation-types, corresponding to the standard propositional attitudes or something like them, and every representation of a given type has the same functional profile. On the permissive view, there are very many roles available - particular representations might play any of a number of different functional roles in a subject's cognitive economy, some of which might look very different from those that fit nicely with the standard propositional attitudes. " [37]

In the paper, he advocates for the permissive view of mental attitudes. As is clear from the passage, the heterogeneity *within* the category of belief supports the characterization of the category itself as porous. This point will be critical for my purposes here.

What is bimagination? Egan presents this mixed state through a metaphor of two libraries. In one library, books are organized into fiction and non-fiction, whereas in the other, the books are tagged with various more complicated tags. In the latter library, users are not as sensitive to the fiction/non-fiction distinction, and may use some of the fiction books occasionally in a way typical of non-fiction, and vice versa - as well as in ways that are not sensitive to the distinction at all. Thus, on his view, belief and imagination are two attitudes held toward the same type of content, a representation of a proposition (or, in the metaphor, a book). The attitudes differ from one another in the use they make of those representations, and in each case, these uses are clusters containing a variety of roles. Bimagination is a case where a representation is used sometimes like imagination and sometimes like belief; since the roles are clusters over time, it's possible to mix and match between some from each cluster at each time or context. He concludes that the Capgras delusion and other monothematic delusions are cases of bimagination.

Egan's view is just designed to treat the intermediate states between belief and imagination, so it has nothing direct to say about the memory cases. While it makes room for intermediate states between all kinds of cognitive attitudes, these memories don't seem to be between belief and some other familiar cognitive state like imagination.

Motivated by cases in which, for instance, I don't already represent the solution to a math problem but I could easily derive it if asked, Robert Audi[4] proposes that the category of dispositions to believe can usefully capture some belief-like states. He distinguishes a dispositional belief, which is stored somewhere in the agent's cognitive system, and a disposition to believe, which is the propensity to form such an encoded representation. Though a disposition to believe that 8.13492 is greater than 8 is not directly stored in the memory system, we can ascribe the disposition in virtue of what is actually stored, and how the agent tends to act. Dispositions to believe can be conditional on the context, and range in strength and persistence. For instance, compare two math problems, one that is far easier to solve.

Audi differs from both Elga and Rayo, and Egan when it comes to the rationality of states in question. Egan takes bimagination to be irrational, and Elga and Rayo do not discuss the rationality of fragmented states directly but treat them as resulting from a limitation in the agent's ability to disperse information. As such, they may not be irrational, but are less ideal than non-fragmented states. Audi, on the other hand, writes:

> "[D]ispositions to believe admit of justification...Note that, at least usually, the grounds of a disposition to believe are internally accessible: S usually can, by reflection, become aware of the crucuial base properties and can often come to know on what ground S is disposed to believe that p. Thus insofar as the grounds are good, one can (normally) use them to justify the disposition or, if they are deficient, to criticize it"[4].

Thus dispositions to believe can be rational or irrational in the ordinary epistemic sense, and without reference to a failure or limitation on the part of the agent.

Audi's view handles distant memories neatly; in those cases, Proust is disposed to believe that he navigated by the church steeple *given that he tastes the madeleine*, just as a glass is disposed to shatter *given that it undergoes a change in temperature*. Likewise, thin memories can be thought of as weaker dispositions to believe, just like my disposition to

believe in the answer to the math problem when the problem is quite difficult. The other cases, like anomalous or recalcitrant memories, seem harder to capture under the heading of dispositions to believe, though of course they can most likely be assimilated into the category of dispositions to think.

Overall, these three accounts are preoccupied with the same problem with which I started this paper; what to do with belief-like states? Elga and Rayo's response broadens the category of belief, whereas Egan and Audi both look for a notion related to belief but outside of it. However, these accounts don't help us with most of the cases of belief-like memory. This is not a reason to reject them per se, but gives us a reason to look further for a theory which will help us understand the relationship between belief-like memories and paradigmatic belief, both descriptively and normatively. Ideally, this will explain why the case of anomalous memory is rational and the Capgras delusion is not.

5.4 The Al-Farabian View

Where should we look for an explanation of the remaining cases? One feature that seems relevant and unexplained in decontextualized and anomalous memories has to do with a deviation in content, rather than attitude. That is, these states differ from belief in *how* they represent the world, not just *under what conditions* they represent the world.

Of course, the attitudes we adopt and the contents towards which our attitudes are directed are not separable. As Gareth Evans put it: "Why does the investigation of the requirements for thoughts and judgement about particular individuals matter? It matters because the concept of thought about an individual is tied to the concept of *understanding* a statement about an individual." (p92)[38] I take this passage to imply that what it means to have a thought with this object as its content should be understood in connection with what we do with such thoughts: namely, we aim to understand statements about their objects. In this section, I'll propose a view of belief-like states that factors these states into attitudes and contents. However, I start with this caveat; while I'll argue that it is explanatorily

advantageous to separate the two questions, they should in the end be thought of as bound together.

Consider this passage from the medieval philosopher Al-Farabi's commentary on Aristotle's *Posterior Analytics*:

> The term "knowing" occurs in a sentence with two meanings - one is "assenting"; the other is "conceptualizing."....Perfect assent is certainty. Perfect conceptualization is to conceptualize something by means of a concise account of what it is in a manner proper to it, because conceptualizing something by means of what signifies it is to define the thing.... Assent may apply both to what is true as well as to what is false. Assent may be certain, it may be approximately certain, it may be the assent that is called"the acquiescence of the soul" with respect to something (which is the one most removed from certainty), and, there is nothing certain whatsoever in false assent. (Kitab al-Burhan)

While introduced in a commentary, this distinction is original to Al-Farabi, and will go on to become a widely-used tool in the classical Islamic tradition[13]. I take Al-Farabi to be describing the space of what we now think of as doxastic states[2]; while he starts the quote by talking about knowledge, it's clear that many of the instances (such as false assent) are not factive, and so would not be covered by current theories of knowledge. This space of doxastic states is distinguished by two parameters, conceptualization and assent:

- Conceptualization: In what way does your representation pick out its object?
- Assent: In what way do you take the representation to be true?

Al-Farabi later writes that assent generally "is for a person to believe concerning a thing about which he makes a judgement that it is, in its existence outside the mind, just as it is believed to be in the mind". How can this description fit a range of attitudes, from certitude

[2]I'll use 'doxastic' to refer to belief and its neighboring representational states

to the acquiescence of the soul? My reading is that assent refers to *how* we take something to be true. This covers not only differences in degree - taking one representation to be more certain than another - but qualitative differences like taking something to be true in a restricted context or for a short time.

This might seem like a permissive reading, given that the paradigm case of assent is certain knowledge. However, there is precedent, at least in the work of Al-Farabi's followers, for taking assent to cover a wide variety of states. For instance, Deborah Black provides the following description of Avicenna's theory of imagination:

> According to Avicenna, all syllogistic arts aim at producing assent to the content of the propositions that form their conclusions, that is, they aim at causing belief in the truth of propositions, according to a variety of degrees of conviction. However, in the case of poetic, and only poetic, syllogisms, Avicenna notes that assent itself is not sought as an end. Rather, the poet utilizes a substitute for assent, namely, the production of an act of the imaginative faculty (*takhyil*) which 'follows the course of what is assented to, due to an impression (*ta'thir*) of it in the soul, this impression in some way taking the place of what causes assent to occur'[13]

While Avicenna extends the concept of assent to states that do not seek out truth, I'll restrict my discussion in this paper to assent which is in some way part of truth-directed inquiry.

Both conceptualization and assent in degrees, and vary beyond just in degree. A precise definition is more conceptualized than a loose sketch, and a deep-seated belief is assented to more fully than a temporary acceptance. However, to expand a little on the text, it's unlikely that these parameters are exhausted by a series of degrees: that is, two loose definitions can be differently conceptualized while still being conceptualized to the same degree. One definition might only include what algae looks like, whereas the other might only include where it can be found. Likewise, I might weakly assent to two propositions,

but in different ways, if I'm disposed to use one in assertion but not other action, and the other in other action but not assertion.

An obstacle for this distinction to be applied to belief and other doxastic states is that we now think of these as *states*. This metaphysical distinction between state and act carries some weight; distinguishing judgment, which is the act that usually results in the formation of a belief, from belief itself allows us to maintain that not all beliefs are formed by an act. Some beliefs, such as the implicit belief that the world does not rest on the back of an elephant, exist in virtue of our other beliefs and have no clear point of origin. Al-Farabi, on the other hand, primarily describes assent and conceptualization as acts, though by classifying them as kinds of knowing, comes closer to describing them as states[3].

I'll pass over the interesting question of why the sharp state/act distinction that we are now familiar with was not emphasized in the works of Al-Farabi to the same degree. Instead, we can make Al-Farabi's view more amenable to contemporary discussion by separating the state from the act which typically brings it about, in both the case of conceptualization and assent. I'll adopt 'representation-formation' and 'assent' to refer to the acts, and 'conceptualization' and 'commitment' to refer to the respective states. Critically, states that involve conceptualization and commitment need not have arisen from acts of representation-formation and assent. Since conceptualization and commitment are merely features of a jointly constituted state towards some content, this move raises significant questions about whether two separate acts typically bring about such a state. I will not attempt to answer these questions here. So we can modify the original sketch to avoid this issue:

- Conceptualization: in what way does your representation present its object?
- Commitment: What behaviors (mental or otherwise) are you committed to in relation to that representation?

[3] As a historical note, Kant also treats assent as fundamental relative to belief[24]

This commitment may be understood as an intrinsic feature of the representation. For instance, by having just the representation of a desire for an apple, I already 'ought' in some weak sense to take steps to acquire an apple. This is not because I have decided to follow through on my desire, or otherwise formed a higher-order attitude toward it, but merely because of what it means to represent the apple as the object of desire.

The Al-Farabian view characterizes a wide range of states, classifying them according to the two parameters. Within this space, we can pick out regions that fit folk psychological categories such as paradigmatic belief, and paradigmatic imagination. But we can also situate the memory phenomenon from the first section; some of these involve less conceptualization than paradigmatic belief, others a different kind of commitment. I'll go through the list in detail in the next section. In addition to these case, the Al-Farabian view neatly captures the difference between the way I believe that string theory is false, and the way that my friend Chip, a philosopher of physics, believes that string theory is false. This difference in conceptualization between my belief and Chip's belief is vast, and has consequences for how we draw inferences and update our belief over time.

This observation is a clue that while the two parameters are detachable, they are by no means independent. That is, it seems that an arbitrary level of conceptualization is compatible with any level of assent. By detachable, I mean that it's in principle possible to firmly hold something you don't represent in very much detail, or to vividly entertain a possibility that you only grudgingly accept. On the other hand, assent may demand increased conceptualization, and increased conceptualization might make assent more reasonable. A stronger view of this connection is exemplified by Descartes, for instance, who takes the liveliness (degree of conceptualization) of a representation to tell us something directly about its accuracy. On Descartes' view, conceptualization makes commitment rational. Alternately, there might be a descriptive connection between commitment and conceptualization, rather than or along with a normative one. Eric Madelbaum[60], for instance, holds that vividly imagining entails belief. In Al-Farabi's terms, this amounts to conceptualiza-

tion entailing commitment.

In summary, the Al-Farabian view treats beliefs, and other mental states such as memory, consideration and guesses, as characterized along the two axes of commitment and conceptualization. This is a complex space, not neatly divisible into kinds separated by gaps. On this view, belief is not a natural or special kind, but merely a term we use for a fuzzily defined region of this space. This picture might seem messy and disorganized; after all, don't we rely on the category of belief for all kinds of philosophical purposes? I'll return to this question in more depth later, but for now it's relevant to note that *secure knowledge* plays the unifying role in Al-Farabi's system.

5.4.1 Conceptualization and Generality

It's now unremarkable to find an entire paper on belief that concerns the 'belief that P', focusing in rich detail on the attitude of belief while avoiding the issue of the content P and how it is represented. However, conceptualization has not entirely dropped out of the contemporary literature on doxastic attitudes. Conceptualization shows up most obviously in discussions of reference and what it takes to have a thought about an object. One relevant debate concerns what is called the generality constraint, roughly the idea that having a thought that some object has some property involves two separable capacities; the capacity to entertain thoughts about a wide range of objects having that property, and to think thoughts about about that object having a wide range of properties. This constraint has been alternately proposed as a requirement on conceptual thought, or even thought in general. For my purposes, it is also a claim about conceptualization, and most interestingly about what conceptualizing one thought tells us about the thinker's ability to conceptualize others.

The generality constraint was first formally proposed by Evans[38], who attributes it to Bertrand Russell's idea that succeeding in thinking 'the cat is fat' implies an ability to differentiate other things from cats, and other properties from fatness. Evans' version of

the constraint seems to treat properties and objects equally, though Dickie[31] argues that a version of the constraint that treats them differently is more true to the original motivation. In particular, she holds that when we think about an ordinary material object, we're in a position to grasp what it would be like for it to have a whole cluster of properties. This type of generality is more fundamental, though not fully general since it only applies to ordinary objects and only invokes the capacity to think about some range of properties as opposed to every property in the thinker's conceptual repertoire. Generality across different objects having the property in question is even more limited; when we think 'a is F', we can't necessarily think 'b is F' but only: 'a is an instance of a kind A, of which other members may or may not be F'.

The generality constraint presents a challenge for the Al-Farabian account. Al-Farabi's view assigns degrees and kinds of conceptualization to individual representations. Implicit in this approach is the idea that the degree and kind of conceptualization can vary widely even among related representations held by the same thinker. The generality constraint, on the other hand, says that in order to think the thought 'a is F', you need to be in a position to conceptualize a wide range of other thoughts. That is, this principle takes the conceptualization of a single thought and infers features about many other potential thoughts. Such an inference should not intuitively be possible if there were indeed a high degree of independence between conceptualization of different representations. Or rather, it should not be possible unless the property being inferred were trivial or not particularly significant for conceptualization overall. A related problem is that the generality constraint treats the ability to think a thought as categorical, as opposed to graded.

I'll consider two responses to this challenge, the second more successful than the first. First, we might maintain that the belief-like representations under discussion are not conceptual, or not fully conceptual, and as such are not subject to the generality constraint. If we take generality to govern conceptual thought, then this would mean ascribing some cases of imperfect conceptualization to non-conceptual thought. Jacob Beck[8], on similar

lines, argues that not all non-general kinds of thought are perceptual, using the example of analog magnitudes. In short, analogue magnitudes are representations of number and amount that are cognized quickly and in a different form than more explicit numerical representations. Since analog magnitudes don't recombine in the way required by generality, Beck concludes that they are non-conceptual. Though the same thing could in principle be said of less-conceptualized representations, it would take these representations into a different cognitive category than more fully conceptualized ones. That is, the division between conceptual and non-conceptual content is typically taken to extend to a division in kinds of processing; for instance, on Christopher Peacocke's[68] model, early non-conceptual perceptual processing feeds into conceptual thought, and so the two kinds of content map on to two stages of thinking. This result would be unfortunate for the Al-Farabian view, since it both imposes a categorical cut-off, and makes this cut-off crucial for the roles of the thoughts in question. By doing so, it violates the separability claim.

More promisingly, Al-Farabi might maintain that despite appearances, the generality constraint is actually graded - and that this gradation reveals that the most reasonable version of the constraint is too trivial to threaten the Al-Farabian account. Elisabeth Camp[19] notes that while advocates of generality limit the constraint to avoid saying that the thinker can entertain thoughts like "Caesar is a prime number", we all manage to understand literary sentences such as "life's but a walking shadow". She argues that the best version of the generality constraint says that if I can entertain thoughts about Caesar, and thoughts about prime number, I should in fact be able to think that Caesar is a prime number. Of course, I will not be able draw very many inferences from this thought, as compared to the thought that 7 is a prime number. But I will be able to get something out of it, as is clear in the case of more usual metaphors. Camp accepts that there is less sense, and less inferential function, to this thought. Her conclusion is that the generality constraint holds in its strongest form, without exceptions - but it prescribes a minimal ability to entertain a thought. On this gloss, the tension between generality and the Al-Farabian view dissolves,

since the Al-Farabian can concede that the wide range of thoughts meets the low standard of generality but within this range there are crucial differences with respect to how much sense the thoughts will make, and whether the thinker can firmly entertain them.

This second response allows the generality constraint and the Al-Farabian view to co-exist. It can be bolstered even further by noting that the generality constraint entails mere capacities to conceptualize, which are of course less demanding than actually conceptualizing or even being disposed to conceptualize. But a takeaway point from this tension is that the Al-Farabian view rests on making some kind of fairly deep functional distinction between representations. Here, this distinction has come into conflict with the idea that we utilize the same capacities across different representations. But this is merely an instance of a broader conflict between a view of doxastic states as somewhat disjointed, and one on which these states are re-combinable units in a flexible mental calculus.

5.5 Returning to the Cases

I'll now come back to the cases of belief-like memory, and demonstrate how the Al-Farabian view handles them. Rather than drag out the discussion of each one, I'll focus mainly on two representative cases: anomalous memories and thin memories. As a reminder, we had:

- **Anomalous memories**: A memory M of event P is treated as an anomaly when the agent singles it out for selective consolidation and attention based in part on the fact that P is surprising, unlikely or odd, and when the agent is at least sometimes not disposed to draw inferences based on M because of P's oddness.

- **Thin memories**: Memories which are extremely information-poor, requiring the details to be filled in with effort and background information (e.g. I have a sketchy memory of a trip I took as a child, but the only way to recognize the event and string the fragments into a narrative is to think very hard about where I could have been,

who would have been there, and so on.)

Anomalous memories involve a tension in conceptualization. The thinker can conceptualize the anomalous event itself quite well. On the other hand, fitting it into her other memories and theories creates a problem; she can't see how they fit together. Of course, she can reject either the memory or some part of her other model of the world - in my example, her conviction in her friend's character. But these cases are marked by the sense that there are other possible resolutions that are not yet able to be conceptualized. Similarly in the Kuhnian case, the scientist can't fit together the anomaly with her background theory, but has a sense that there is something to be done with the anomaly in the future, or that it's a hint that something could be going on that she is not yet aware of. These descriptions over-intellectualize the process, which is most often entirely opaque to the thinker, but this is just to make the case fully explicit. Anomalous memories feature a conceptualization that is vivid internally to the representation of the event, yet lacking (or seemingly lacking) in relation to the broader representation. This challenge in conceptualization is not impossible to resolve; there is the obvious solution of rejecting either part of the memory or part of the broader representation. And yet, our thinker does not take this solution. Instead, she undertakes a commitment to resolve the problem later, and in the meantime, be on the alert for relevant evidence and take advantage of time to think through other possibilities. That is, this degree of conceptualization is compatible with the more familiar commitment to resolve inconsistency now, but this commitment is absent in the case of anomalous memories.

Fragile memories are similar, especially when the recalcitrance stems from a contradiction with other beliefs or memories; the main difference is that the implicit standing commitment to taking the content to be at all accurate, a far weaker commitment than in the anomalous memory case, is rejected upon retrieval. Recalcitrant memories may be perfectly conceptualized, and we have a strong commitment to taking them to be true in a wide variety of ways except for one crucial one - we don't take them to be faithful.

De-contextualized memories are so poorly conceptualized that they could not possible be committed to, since they fail to describe the sort of thing that could be true or false.

Thin memories are more interesting, because they exploit the connection between commitment and conceptualization in a similar way to anomalous memories. These memories, like de-contextualized memories, are loosely conceptualized. On the other hand, they differ in that we have both the ability as well as the propensity to construct a fuller conceptualization. The ability to do so reflects a combination of the kind of conceptualization, and the abilities and knowledge of the agent at the point of reconstruction. The propensity, on the other hand, will presumably depend on the importance of the reconstruction to the thinker relative to their other priorities, in just the same way that I can make many more inferences from what I already know that I will actually make. This propensity may also be partly explained by the memory itself; we could say that there is a default propensity to reconstruct a memory, since the memory was originally formed with some kind of commitment to its content, and the only way to realize this commitment would be to first reconstruct the content. This relationship - where a partial conceptualization indicates an unfulfilled commitment - can be altered and refined as the memory is reconstructed. As the thinker begins to understand what the partially-remembered event actually was, she might realize that what seemed to be a memory was actually a confused imagining, and thus no further commitment would be called for, as in the case of fragile memories. On the other hand, she may start to see that this event has crucial implications that she was ignorant of, and so commit more strongly to the content and be more invested in reconstructing it fully. This back-and-forth between reconstructing the content and evaluating it and its consequences is an instance of the familiar observation in cognitive science: effective search involves an integrated loop between retrieval and evaluation.

5.5.1 Comparing the Taxonomies

It should not be particularly surprising or convincing that the Al-Farabian view can accommodate these examples; after all, it has far more flexibility than either restrictive theories of belief or the more flexible theories of belief-like states that I discussed above. I will briefly compare how these cases can be treated under the accounts in section 5.3, with the caveat that this is not meant to be an argument against those other accounts. Since they have distinct aims from the view I'm developing, it would be odd to take the failure to accommodate all kinds of memory cases as a mark against the theories.

On the fragmentation view, distant memories can be redescribed as beliefs given conditions. However, many of these cases, such as thin memories, involve bilateral connections between what is recalled and what elicits it; as we recall some of the details, we will enhance the memory which in turn leads us to uncover further relevant recorded details, as I've just described in the Al-Farabian framework. That is, no table of eliciting conditions could describe these cases, since they involve a loop between eliciting conditions and stored information that is iterated and moves in both directions. Further, how anomalous memories and recalcitrant memories differ from paradigmatic belief can't be diagnosed the fragmentation view, since they do not differ in when they are elicited, but rather in how they are used and updated. These objections derive from the fact that fragmentation plays with the attitudes toward propositional contents, but does not give us tools to discuss differences in the the representation. But we need differences in representations of the same proposition in order to differentiate these memories, and to discuss thin and de-contextualized memories that have genuinely, and obviously, non-standard contents.

On Audi's dispositionalist view, we can in principle capture many of these phenomena as some kind of disposition. For instance, anomalous memories are dispositions to believe in some contexts, but not in the inferential context. Likewise, this account makes quite good sense of thin memories, since it can appeal to the cognitive effort involved in filling in the picture. De-contextualized and fragile memories might be more of a challenge, since

they do not result in belief.

However, even were they to accommodate these belief-like phenomena, these accounts all have a hard time separating defective states like the Capgras delusion from two other categories: (a) states that are rational responses to information processing limitations, (b) states that are rational for any kind of creature. But many memory phenomena fit into (a) and possibly even (b). To see the contrast between the picture I am outlining and these alternative strategies, consider, for instance, Egan's claim about the normative superiority of non in-between states: "The answer, very briefly, is that agents whose representations tend to have something like the stereotypical roles will tend to do much better than those whose representations tend to have intermediate, mix-and-match roles. We shouldn't expect to see the peculiar hybrid roles all over the place because they're pretty maladaptive." [37]. That is, neither theory tells us about why these kinds of belief-like states might be part of a well-functioning epistemic life.

On the other hand, the way I've described the cases on the Al-Farabian view is the start of a rational explanation. For instance, the diagnosis of thin memories shows how the degree of conceptualization starts the thinker off with a prima facie reason to reconstruct, given that it suggests a lost commitment. The description of how the process unfolds as a loop between improving the conceptualization and evaluating what commitments are and should be in place allows us to say that some instances of the process are more reasonable than others, in the sense that they are more responsive to the commitments the thinker has. This process may be entirely inaccessible, but it follows the kind of reasoning that if accessible, would be evaluable. Whether this is enough for true rational evaluation is controversial. But by appealing to a notion of commitment that is an extension of the one we are already familiar with in paradigmatic belief, the Al-Farabian is in a good position to make the case for the rational evaluability of cognitive commitment in general. Unfortunately, this is beyond the scope of the present chapter. My aim here is only to show that this medieval theory is a promising route to understanding the relationship between belief

and neighboring states in memory and elsewhere in cognition.

5.6 You Only Are What You Believe?

Writing this paper, I was reminded of a folk song written by Phil Ochs, in which he paints a picture of an absurdist anti-war strategy, that of giving up on sincere moral protest and instead declaring, and believing, that the war is over. The song ends:

The gypsy fortune teller told me that we'd been deceived
You only are what you believe
I believe the war is over
It's over, it's over

I mention this song as a reminder that belief has a central place in the way we think of ourselves now. The protester who believes that the war is over does not just see the world in a certain way but expresses something radical, and even might bring about a change in the world. The song about imagining the war to be over, or guessing that it's over, would not make the same point. [4]

Even if you found the arguments of the preceding section persuasive, they leave unsolved a serious issue. If we move from the contemporary picture of belief as the canonical epistemic attitude, what would be lost? That is, belief has a special role to play in our broader epistemology, and even ethics and metaphysics, and so by treating the category of belief as a fairly arbitrary region of a more complex range of attitudes, the Al-Farabian view may lead to untenable theoretical consequences.

In other words, belief has a special role (or roles) in our epistemology, and in connecting epistemology to other areas. What are these special roles? While much has been said on this question, I'll focus on just three:

[4]This intuitive significance we associate with belief can also be seen through the use of 'belief in'. I owe this point to Peter Railton.

1. **primary agential role**: Believing is an exercise of agency, which explains why we can be held accountable for our beliefs (among other things)

2. **secondary agential role**: Belief guides our intentions, which explains in what way we can be held accountable for our actions and their consequences.

3. **epistemic role**: Belief, when successful, becomes knowledge, which is the aim of our epistemic lives.

The primary agential role is notably discussed by William Clifford[27], in *The Ethics of Belief*; Clifford distinguishes belief from attitudes short of belief such as suspicion or acceptance by the fact that we are distinctively ethically responsible for our beliefs. Others rely on our beliefs, and we build our epistemic character from our practice of believing. Clifford's picture emphasizing the accountability we have for our beliefs themselves. The other component of the primary agential role is about how beliefs are formed; in a recent paper, Matthew Boyle[16] argues that belief is distinctively agential, even though it is not an occurrent act, but a state sustained over time. Boyle and Clifford take up a very strong position, that belief itself is ethically evaluable and an exercise in agency. More commonly accepted is the weaker secondary agential role, in which belief guides intentional action, and since intentional action itself is the subject of ethical (and practical) evaluation, this places a derivative ethical or practical significance on belief. The epistemic role is implicit in the literature of knowledge as belief plus some other conditions, though it has been challenged by Williamson[95] and other proponents of knowledge-first epistemology.

On the Al-Farabian view, belief is one of many similar and interrelated attitudes that play a range of related roles. This creates a prima facie tension with both the agential and epistemic pictures in the following way. The agential view, most vividly in the primary agential role as expressed by Clifford and Boyle, elevates belief *categorically* - all and only belief has the relevant agential and ethical properties. Combining this with the Al-Farabian account generates two explanatory problems: (a) why would the normative properties im-

pose a sharp cut-off around a vague and fuzzy descriptive category? and (b) why would only belief, as opposed to other combinations of assent and conceptualization bear these normative properties?

The secondary agential role of belief generates an additional, more practical problem. Consider our starting case of an anomalous memory: this state involved a deviation from the norms of belief that require following through on inferring the probable consequences of a belief when salient. This deviation results in acting differently than one would if one believed the memory in a more paradigmatic sense. The secondary agential role posits that belief guides intention, so this kind of case is troubling because it implies a lack of guidance - or leniency in guidance. After all, this attitude, and many of the other types of belief-like memory, by definition involve committed deviation from the norms regulating belief, and thereby a potential deviation from guidance by reason itself. There are two ways to treat this deviation: (a) as unguided, or (b) as guided by another belief or set of beliefs in combination with desire and so on. Both options lead to further problems: (a) implies that either these cases are defective, or the role of guidance is significantly more limited that we might have thought, and (b) generates a difficult, and perhaps, intractable task of regulating these varied commitments.

I cannot fully resolve these tensions here, but I will point to two avenues that might be fruitful in re-configuring these undeniably important elements of our mental lives, agency and epistemic value, to fit the pluralistic Al-Farabian account.

One attractive solution to the agency challenge is to argue that all commitments, not just belief, are exercises in agency for the primary role, or foundations of agency for the secondary role. One route uses the notion of commitment itself as the locus of agency. After all, the notion of commitment is itself agential - a person makes a commitment, and is bound by ethical norms as a result. Searle[79] controversially holds that commitments such as promises are descriptive acts that generate normative consequences. This machinery might be re-purposed for the case of cognitive commitments; just like promises, dox-

astic states have ethical evaluability because of the descriptive features of the act combined with background conditions. However, in Searle's case (which is itself controversial), the normative consequences only follow because of the *institution* of promising. So there are at least two significant issues to be addressed. First, it's unclear whether anything like an institution could exist in the case of assent - though there might be some kind of substitute in a Millikan-like theory of cognitive function, or in metacognition. Further, this structure does not explain how we can evaluate making commitments in the first place, but only gives us grounds for evaluating whether existing commitments should be satisfied.

In regards to the epistemic role, the tradition of classical Islamic epistemology might have an answer. Al-Farabi's own view of knowledge used the ingredients of conceptualization and assent to define something incredibly demanding; knowledge is firm and non-accidental assent to something necessarily true[12][5]. This is a natural fit with Plato's contrast between knowledge and opinion in the Republic. There, opinion (doxa) takes as its object fragile, contingent facts - such as the color of a chair - whereas knowledge is of necessary truths. For both Plato and Al-Farabi, knowledge is not only demanding in terms of the relationship between the knower and the thing known, but also requires that the thing known be of a certain kind beyond the kind of thing we normally believe or assent. On this picture, the epistemic role of belief defined above needs to be redefined anyways; only a very small subset of beliefs are even about the right kind of thing for knowledge. The other beliefs, about contingent matters, surely play some kind of epistemic role, though not the one specified above. Whether this is an instrumental role, oriented towards coming to know about other things, or something else is beyond the scope of this paper. But these older epistemological projects suggest a way to locate the epistemic role of the wider variety of doxastic states presented in this chapter. That is, knowledge is on many views extremely demanding, so demanding that for most of the questions that we consider, we

[5]Black actually includes belief as one component in this account from Al-Farabi's *Conditions of Certitude*, though she notes that Al-Farabi treats it as a common-sense linguistic category, and that knowledge is instead defined in terms of assent in the commentary on the *Prior Analytics*, which I used to set up the view earlier.

should already be in a position to see that our answer in the best case will not amount to knowledge. If any of these everyday cases amount to belief, and if belief is distinguished by a relationship with knowledge, then we already need to think of this relationship in more indirect terms than 'I believe P *now* in order for my belief to amount to knowledge *now*'. Thus, moving from belief to other states in the space of assent and conceptualization does not involve any kind of substantial weakening the relationship with knowledge. Instead, what we already needed on the belief picture is made more obvious on the Al-Farabian view: a theory to explain the epistemic value of states that are points on the way to knowledge but not themselves candidates for knowledge [6]

[6]The most familiar contemporary form of such a theory is epistemic consequentialism, but I will end by sketching a different version of this kind of theory as developed by Al-Farabi's most notable follower, Ibn Sina (Avicenna). Roughly, Ibn Sina draws on Aristotle's theory of potentiality to describe the process of coming to know as actualization of potential. This potential adheres in the active intellect, which Ibn Sina alternately describes as the heavenly intellect. This intellect, the thing that is actualized, the standard for knowledge, doesn't obtain merely internally as a goal, but has its own existence. It is what is actualized when any knower comes to know, not a personal faculty of anyone in particular. On his analogy, a child has some potential for writing by merely existing, but as he learns hand-coordination and words, he attains a higher stage of potential. Further stages are actualized as he learns to write, and finally writes so often that it becomes a habit. Thus, on Ibn Sina's view, states in the space of assent and conceptualization partially actualize the active intellect, or knowledge itself, without being themselves candidates for knowledge. This notion of potential - where the potential is a sort of thing that is actualized but already always existed in some sense - is markedly distinct from most common-sense contemporary understandings.

CHAPTER VI

Conclusion

This dissertation has investigated what it takes to learn about our environment over time, and given the complex relationships between our mental representations, mental and physical behaviors, and the possibility of acquiring more evidence. I hope to have shown that memory has a part to play in this story that is neither simple nor derivative of other cognitive systems. But I'd like to end on a note of caution.

This project endorses a kind of holism, where rational change over time cannot be reduced to a series of independently rational transitions between time-slices, and representations over interrelated facts cannot be reduced to a series of independent 'beliefs that P'. Holism in this sense is contrasted with atomism, the view that more complex processes should be understood as composites of smaller parts that can be analyzed for the most part on their own. While I ultimately think a holistic view of the epistemology of learning and memory is our best option, holism has some serious dangers.

Any kind of holism, by denying atomistic reduction, is at least *prima facie* less simple than atomism. Holism can also be mystical and un-explanatory by withholding the possibility of reductive explanation. And worse, it can be cheap. After all, it is easier to convince people that things are more complicated than they appear than that they are simpler - in a sense, simplicity and reduction are theoretical achievements, whereas holism is a step away from these achievements. We might also worry that holism is hard to escape; once things

have been muddied, interdependent and complicated, it becomes more difficult to falsify the theory.

So the bar for adopting a holistic theory should be very high. To my mind, the case for the kind of diachronic holism I've advanced rests on two features. There should be a compensatory explanatory advantage, and the inquiry should be responsive to a wider variety of potentially challenging evidence. Or in simpler terms, the only reason to go in for holism despite these flaws is if it gives us insights that were otherwise unreachable, and can be falsified by a range of available data so that we do not accept it too easily.

On the former point, I hope to have shown that this holistic picture gives us insight into previously unexplored *dynamic structures*, not just the ability to describe some new cases. For instance, the exploration/exploitation trade-off in Chapter III, or the conceptualization/commitment model of doxastic states in Chapter V are steps towards holism that are also supposed to provide a way to schematize and predict. As opposed to a holism that replaces atomist reduction with the claim that the phenomenon in question is deeply particular and impossible to factor, this kind of holism aspires to introduce new explanatory machinery.

On the latter point, this project built rational theories inspired by empirical observations, and aimed to hold these theories up to empirical scrutiny. As opposed to more theoretical philosophical models of memory, my picture is subject to falsification; since I motivated many of the steps by appealing to how things work in successful agents, finding out that things work quite differently than I have described, or that these agents are not so successful after all, would make my view far less plausible. So while holism has a genuine problem with falsification, if holism is part of what makes this project empirically applicable, holism also makes this project subject to empirical falsification and constraint.

In conclusion, I have shown how a diachronic picture of rationality is both genuinely different from a synchronic one, and makes contact with a wide and exciting range of empirical findings. This dissertation is an initial step towards a holistic theory of learning

and memory that uncovers rational structures behind complex behaviors.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Peter Adamson. *Classical Arabic philosophy: sources and reception*, volume 11. Warburg Institute, 2007.

[2] Arif Ahmed and Bernhard Salow. Dont look now. *The British Journal for the Philosophy of Science*, 2017.

[3] John R Anderson and Robert Milson. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703, 1989.

[4] Robert Audi. Dispositional beliefs and dispositions to believe. *Noûs*, 28(4):419–434, 1994.

[5] Robert Audi. Doxastic voluntarism and the ethics of belief. *Knowledge, truth, and duty*, pages 93–111, 2001.

[6] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

[7] David James Barnett. Is memory merely testimony from one's former self? *Philosophical Review*, 124(3):353–392, 2015.

[8] Jacob Beck. The generality constraint and the structure of thought. *Mind*, 121(483):563–600, 2012.

[9] Gordon Belot. Bayesian orgulity. *Philosophy of Science*, 80(4):483–503, 2013.

[10] Daniel Bendor and Matthew A Wilson. Biasing the content of hippocampal replay during sleep. *Nature neuroscience*, 15(10):1439, 2012.

[11] Sven Bernecker. *Memory: A Philosophical Study*. Oxford University Press, 2010.

[12] Deborah L Black. Knowledge (ilm) and certitude (yaqīn) in al-fārābīs epistemology. *Arabic Sciences and Philosophy*, 16(1):11–45, 2006.

[13] Deborah L Black. Certitude, justification, and the principles of knowledge in avicennas epistemology. *Interpreting Avicenna: Critical Essays*, page 120, 2013.

[14] Jennifer Urbano Blackford, Joshua W Buckholtz, Suzanne N Avery, and David H Zald. A unique role for the human amygdala in novelty detection. *Neuroimage*, 50(3):1188–1193, 2010.

[15] Barak Blumenfeld, Son Preminger, Dov Sagi, and Misha Tsodyks. Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52(2):383–394, 2006.

[16] Matthew Boyle. 'making up your mind' and the activity of reason. *Philosopher's Imprint*, 11(17), 2011.

[17] Tyler Burge. Interlocution, perception, and memory. *Philosophical Studies*, 86(1):21–47, 1997.

[18] György Buzsáki and Edvard I Moser. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature neuroscience*, 16(2):130–138, 2013.

[19] Elisabeth Camp. The generality constraint and categorial restrictions. *The philosophical quarterly*, 54(215):209–231, 2004.

[20] J. Campbell. *Reference and Consciousness*. Oxford cognitive science series. Clarendon Press, 2002.

[21] Jennifer Rose Carr. Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research*, 95(3):511–534, 2017.

[22] David J. Chalmers. The representational character of experience. In Brian Leiter, editor, *The Future for Philosophy*, pages 153–181. 2004.

[23] Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2004.

[24] Andrew Chignell. Belief in kant. *The Philosophical Review*, 116(3):323–360, 2007.

[25] Marcus Tullius Cicero, James M May, and Jakob Wisse. *Cicero on the ideal orator (De Oratore)*. Oxford University Press, USA, 2001.

[26] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03):181–204, 2013.

[27] William Kingdon Clifford. *The ethics of belief and other essays*. Prometheus Books, 1999.

[28] Laura Lee Colgin, Edvard I Moser, and May-Britt Moser. Understanding memory through hippocampal remapping. *Trends in neurosciences*, 31(9):469–477, 2008.

[29] Max Coltheart, Robyn Langdon, and Ryan McKay. Schizophrenia and monothematic delusions. *Schizophrenia Bulletin*, 33(3):642–647, 2007.

[30] Tim Dalgleish, Lauren Navrady, Elinor Bird, Emma Hill, Barnaby D Dunn, and Ann-Marie Golden. Method-of-loci as a mnemonic device to facilitate access to self-affirming personal memories for individuals with depression. *Clinical Psychological Science*, 1(2):156–162, 2013.

[31] Imogen Dickie. The generality of particular thought. *The philosophical quarterly*, 60(240):508–531, 2010.

[32] Susanne Diekelmann, Ines Wilhelm, and Jan Born. The whats and whens of sleep-dependent memory consolidation. *Sleep medicine reviews*, 13(5):309–321, 2009.

[33] Stephane Dissel, Krishna Melnattur, and Paul J Shaw. Sleep, performance, and memory in flies. *Current sleep medicine reports*, 1(1):47–54, 2015.

[34] Jeffrey M Donlea, Matthew S Thimgan, Yasuko Suzuki, Laura Gottschalk, and Paul J Shaw. Inducing sleep by remote control facilitates memory consolidation in drosophila. *Science*, 332(6037):1571–1576, 2011.

[35] Spyridon Drosopoulos, Eike Windau, Ullrich Wagner, and Jan Born. Sleep enforces the temporal order in memory. *PLoS One*, 2(4):e376, 2007.

[36] NMJ Edelstyn and Femi Oyebode. A review of the phenomenology and cognitive neuropsychological origins of the capgras syndrome. *International Journal of Geriatric Psychiatry*, 14(1):48–59, 1999.

[37] Andy Egan. Imagination, delusion, and self-deception. *Delusions, self-deception, and affective influences on beliefformation*, 2009.

[38] Gareth Evans. *The Varieties of Reference*. Oxford University Press, Oxford, UK, 1996.

[39] Edward Fredkin. Trie memory. *Communications of the ACM*, 3(9):490–499, 1960.

[40] Jane Friedman. Suspended judgment. *Philosophical studies*, 162(2):165–181, 2013.

[41] Irving J Good. On the principle of total evidence. *The British Journal for the Philosophy of Science*, 17(4):319–321, 1967.

[42] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[43] Hilary Greaves. Epistemic decision theory. *Mind*, 122(488):915–952, 2013.

[44] H Griessenberger, K Hoedlmoser, DPJ Heib, J Lechinger, W Klimesch, and M Schabus. Consolidation of temporal order in episodic memories. *Biological psychology*, 91(1):150–155, 2012.

[45] Alden L Gross, Jason Brandt, Karen Bandeen-Roche, Michelle C Carlson, Elizabeth A Stuart, Michael Marsiske, and George W Rebok. Do older adults use the method of loci? results from the active study. *Experimental aging research*, 40(2):140–163, 2014.

[46] Pernille Hemmer and Mark Steyvers. A bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1):189–202, 2009.

[47] Christopher Hitchcock and Elliott Sober. Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science*, 55(1):1–34, 2004.

[48] Simon M Huttegger. Bayesian convergence to the truth and the metaphysics of possible worlds. *Philosophy of Science*, 82(4):587–601, 2015.

[49] Simon M Huttegger. *The Probabilistic Foundations of Rational Learning*. Cambridge University Press, 2017.

[50] Tommi Jaakkola and Hava Siegelmann. Active information retrieval. 2001.

[51] William James. *The will to believe and other essays in popular philosophy*, volume 6. Harvard University Press, 1979.

[52] Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. 2015.

[53] Philip Kitcher. Theories, theorists and theoretical change. *The Philosophical Review*, 87(4):519–547, 1978.

[54] John R Krebs, Alejandro Kacelnik, and Peter Taylor. Test of optimal sampling by foraging great tits. *Nature*, 275(5675):27–31, 1978.

[55] Thomas S Kuhn. *The Structure of Scientific Revolutions, 2nd enl. ed.* University of Chicago Press, 1970.

[56] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[57] Elizabeth F Loftus. *Eyewitness testimony*. Harvard University Press, 1996.

[58] Aleksandr Romanovich Luria. *The mind of a mnemonist: A little book about a vast memory*. Harvard University Press, 1968.

[59] Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, pages 121–151. Springer, 2008.

[60] Eric Mandelbaum. Thinking is believing. *Inquiry*, 57(1):55–96, 2014.

[61] Jutta S Mayer, Jejoong Kim, and Sohee Park. Failure to benefit from target novelty during encoding contributes to working memory deficits in schizophrenia. *Cognitive neuropsychiatry*, 19(3):268–279, 2014.

[62] Sam McKenzie and Howard Eichenbaum. Consolidation and reconsolidation: two lives of memories? *Neuron*, 71(2):224–233, 2011.

[63] Kourken Michaelian. The epistemology of forgetting. *Erkenntnis*, 74(3):399–424, 2011.

[64] Kourken Michaelian. Generative memory. *Philosophical Psychology*, 24(3):323–342, 2011.

[65] Kourken Michaelian. *Mental time travel: episodic memory and our knowledge of the personal past*. MIT Press, 2016.

[66] Vishnu P Murty and R Alison Adcock. Enriched encoding: reward motivation organizes cortical networks for hippocampal detection of unexpected events. *Cerebral Cortex*, 24(8):2160–2168, 2013.

[67] Lars Nyberg. Any novelty in hippocampal formation and memory? *Current opinion in neurology*, 18(4):424–428, 2005.

[68] C Peacocke. Does perception have a nonconceptual content? *The Journal of Philosophy*, 98:239–264, 2001.

[69] Gina R Poe, Christine M Walsh, and Theresa E Bjorness. Cognitive neuroscience of sleep. *Progress in brain research*, 185:1, 2010.

[70] William H Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392, 2009.

[71] Travis Proulx, Michael Inzlicht, and Eddie Harmon-Jones. Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in cognitive sciences*, 16(5):285–291, 2012.

[72] Peter Railton. Truth, reason, and the regulation of belief. *Philosophical Issues*, 5:71–93, 1994.

[73] V. S. Ramachandran. Consciousness and body image: lessons from phantom limbs, capgras syndrome and pain asymbolia. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 353(1377):1851–1859, 1998.

[74] Charan Ranganath and Gregor Rainer. Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, 4(3):193–202, 2003.

[75] Mauricio Rangel-Gomez, Clayton Hickey, Therese van Amelsvoort, Pierre Bet, and Martijn Meeter. The detection of novelty relies on dopaminergic signaling: evidence from apomorphine's impact on the novelty n2. *PLoS one*, 8(6):e66469, 2013.

[76] Geraldine Rauchs, Dorothee Feyers, Brigitte Landeau, Christine Bastin, Andre Luxen, Pierre Maquet, and Fabienne Collette. Sleep contributes to the strengthening of some memories over others, depending on hippocampal activity at learning. *The Journal of Neuroscience*, 31(7):2563–2568, 2011.

[77] Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

[78] Nina Rouhani, Kenneth A Norman, and Yael Niv. Dissociable effects of surprising rewards on learning and memory. *bioRxiv*, page 111070, 2017.

[79] John R Searle. How to derive "ought" from "is". *The Philosophical Review*, 73(1):43–58, 1964.

[80] Wilfrid Sellars et al. Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, 1(19):253–329, 1956.

[81] Thomas D Senor. Preserving preservationism: A reply to lackey. *Philosophy and Phenomenological Research*, 74(1):199–208, 2007.

[82] Nishi Shah and J David Velleman. Doxastic deliberation. *The Philosophical Review*, pages 497–534, 2005.

[83] Stephen F Signer. Capgras syndrome: the delusion of substitution. *J Clin Psychiatry*, 48(4):147–150, 1987.

[84] S. Singh, R.L. Lewis, A.G. Barto, and J. Sorg. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, June 2010.

[85] Chandra Sripada. Imaginative guidance: A mind forever wandering. *Homo Prospectus*, page 103, 2016.

[86] Gisela Striker. *Aristotle's Prior Analytics Book I: Translated with an Introduction and Commentary*. Oxford University Press, 2009.

[87] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[88] J.D. Sweatt. *Mechanisms of Memory*. Elsevier Science, 2009.

[89] Stefano Tracà and Cynthia Rudin. Regulating greed over time. *arXiv preprint arXiv:1505.05629*, 2015.

[90] Dorothy Tse, Rosamund F Langston, Masaki Kakeyama, Ingrid Bethus, Patrick A Spooner, Emma R Wood, Menno P Witter, and Richard GM Morris. Schemas and memory consolidation. *Science*, 316(5821):76–82, 2007.

[91] Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980.

[92] Marlieke TR van Kesteren, Dirk J Ruiter, Guillén Fernández, and Richard N Henson. How schema and novelty augment memory formation. *Trends in neurosciences*, 35(4):211–219, 2012.

[93] J. David Velleman. Epistemic freedom. *Pacific Philosophical Quarterly*, 70(1):73–97, 1989.

[94] Matthew P Walker and Robert Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44(1):121–133, 2004.

[95] Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.

[96] Yuni Xia, Reynold Cheng, Sunil Prabhakar, Shan Lei, and Rahul Shah. Indexing continuously changing data with mean-variance tree. *International journal of high performance computing and networking*, 5(4):263–272, 2008.

[97] Seda Yilmaz, Colleen M Seifert, and R Gonzalez. Cognitive heuristics in design: Instructional strategies to increase creativity in idea generation. *AI EDAM*, 24(3):335–355, 2010.

[98] Weiwei Zhang and Steven J Luck. Sudden death and gradual decay in visual working memory. *Psychological science*, 20(4):423–428, 2009.

[99] Kevin JS Zollman. The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587, 2007.