Approaching Epistemology Through Memory

A Thesis
Presented to
The Division of Philosophy, Religion, Psychology, and Linguistics
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Alexander Nord

May 2014

Approved for the Division
(Philosophy)

_____

Paul Hovda

Acknowledgments

Foremost, I would like to thank Paul Hovda—without his guidance I certainly could not have produced this thesis (or a thesis identical to any element in the set of all equally good or better possible theses). I would like to thank my family for graciously supporting me in uncountably many ways during the past 21.5 years. I would like to thank in order (though I will not divulge what order) Justin Katz and Dylan Whitlow for some of the things that they have done, as well as some of the other things that they have done, but not all of the things that they have done—without certain of the things you did my sanity would not have survived this year intact. Finally, and least of all, I feel obligated to acknowledge Epsilon as the worst thing ever.

Table of Contents

Abstract

Almost every exercise of human cognition relies on the use of memory, but as a subject of philosophical interest memory receives comparatively little attention. My aim in writing this thesis is to demonstrate that memory is a philosophically substantial topic by arguing that insofar as we cannot adequately understand what an agent believes without developing an adequate theory of memory, considerations of what memory is and how it functions are epistemologically indispensable.

I begin by discussing David Lewis's claim that it is formally beneficial to model cognition around multiple fragmented belief structures, instead of a single belief set. After examining how prior attempts at developing fragmentation-based models of belief fall victim to psychological plausibility concerns, I propose an alternative model, the Single Fragment Model, which serves as a rough framework of a psychologically plausible approach to formally modeling the organization of our beliefs.

In my second chapter I investigate whether justifications can be preserved through memory in the way that Tyler Burge's theory of preservative memory describes. I begin by discussing a concern raised by David Christensen and Hilary Kornblith that because memory is a reconstructive mechanism, memory-beliefs are not preserved with their original counterparts' justifications. In order to understand how we should think of memory as a reconstructive process I investigate the notion of memory traces, as described by Mohan Matthen. Using the theory of memory traces, I develop a principle by which one can determine the degree to which justification is transmitted through the memorial process, and use this "Memory Justification Principle" to show that in certain cases memory-beliefs may be thought of as been produced by the sort of preservative memory that Burge defends. I also note that memory trace theory supports the Single Fragment Model of cognition described in the first chapter.

My final chapter considers an argument put forward by Jennifer Lackey that the Preservation View of Memory, which is a necessary consequence of my Memory Justification Principle, is false. Lackey argues against the Preservation View of Memory through cases which employ normative defeaters. After investigating the notion of normative defeaters I determine that Lackey succeeds in defeating the Preservation View of Memory, and conclude by adjusting the Memory Justification Principle so as to make it compatible with the results of Lackey's argument.

The work of the philosopher consists in marshalling recollections for a particular purpose.

– Ludwig Wittgenstein, *Philosophical Investigations* (§127)

Introduction

Almost every exercise of human cognition relies on the use of memory, but as a subject of philosophical interest memory receives comparatively little attention. One might argue that this is because memory is too transparent of a topic to be interesting in its own right—memory simply presents us with information with which we have previously engaged, and whatever problems memory presents will ultimately reduce to problems about how we acquired that information in the first place. One might argue that the subject of memory is neglected in philosophy because it is too opaque—it functions entirely within the inscrutable processes of the mind (or brain) and leaves little to talk about. In any case, memory, as a philosophically rich subject, is rarely given the credit that it deserves. My aim in writing this thesis is to demonstrate that memory is a philosophically substantial topic by arguing that insofar as we cannot adequately understand what an agent believes without developing an adequate theory of memory, considerations of what memory is and how it functions are epistemologically indispensable.

I begin the first chapter of my thesis by discussing a claim, put forward by David Lewis, that instead of modeling agents as though they possess single monolithic sets of beliefs, we should formalize our representations of agents' doxastic structures by thinking of beliefs as somehow organized into a number of distinct fragmentary belief sets, which are presumably organized around particular themes. Under the monolithic approach to belief modeling, we attribute to agents the logical closure of their explicit beliefs, such that we would attribute a belief that $P \wedge Q$ to an agent who explicitly believes P and explicitly believes Q. This creates problems, however, when an agent's belief set allows us to attribute to that agent both P and $\neg P$, for this allows us to attribute a belief in $P \wedge \neg P$ to the agent, which then commits the agent to all expressible propositions, *ex falso quodlibet*. The benefits of fragmentation are that it gives us a more realistic illustration of the mind's structure than the traditional single belief set picture, and also provides us with a theoretical means for preventing the logical closure of inconsistent beliefs from committing their believer to all expressible propositions—so long as inconsistent beliefs are contained in separate fragments, fragmentation theory prevents the believing agent from being committed to the logical closure of those inconsistent beliefs, and thus protects the agents it models from becoming committed to all expressible propositions, *ex falso quodlibet*. I then turn to examine Andy Egan's conception of belief fragmentation and the *B*-Structures Model developed by Samir Chopra and Rohit Parikh, but determine that both of these fragmentation models fail to be psychologically plausible in virtue of

their ontologically robust groupings of beliefs into fragments. In response to these shortcomings I propose an alternative model, the Single Fragment Model. The distinguishing feature of the Single Fragment Model is that it treats each of an agent's explicit beliefs as an individual fragment, but makes use of the notion of an "active" fragment as a variable fragment whose constitution depends on which individual beliefs are "active" within the agent's present cognitive context. The primary benefit of the Single Fragment Model is that it enjoys all of the formal benefits of belief fragmentation (*e.g.*, avoiding the problems that arise from agents harboring logically inconsistent beliefs), while also enjoying a high degree of psychological plausibility, in virtue of being formally analogous to memory. We see that basing our formal models of belief on memory mechanisms gives us an optimally intuitive account of agents' belief structures.

In the second chapter I transition into a discussion of memory itself, rather than the highly abstract activation and deactivation processes used by the Single Fragment Model. I begin this chapter by introducing Tyler Burge's notion of preservative memory, which may be roughly thought of as any memory mechanism which preserves mental state contents along with their original justificatory supports, and which does not provide any further justificatory support or hinderance for the remembered contents. After presenting the Burge's theory of preservative memory, I consider an objection raised by David Christensen and Hilary Kornblith, who argue that contemporary cognitive psychology characterizes memory as a "reconstructive" process, and that preservative memory is incompatible with this fact. After demonstrating that two possible interpretations of what Christensen and Kornblith might mean by calling memory a reconstructive process both fail to defeat Burge's theory of preservative memory, I consider a third account of what it means for memory to be reconstructive. This third approach to understanding memory as reconstructive is the theory of memory traces discussed by Mohan Matthen, which posits the existence of memory traces as non-contentful mental objects which perform the functions of storing and being able to reproduce particular mental states (with variable accuracy). The essential idea behind memory traces is that they are "sculpted" with respect to particular mental states, and can (in normal circumstances) be stimulated to produce a mental state with similar content, albeit with the addition of a minor amount of signature content. I then examine in detail how we are able to understand the content produced by activated memory traces, and investigate the inferential patterns necessary to arrive at memory-based beliefs that are identical (in terms of content) to their original counterparts. Following my discussion of memory trace theory, I present the Memory Justification Principle as an account of how justification is transmitted through memory within the reconstructive picture of memory developed during my discussion of memory trace theory. The Memory Justification Principle reveals that in certain cases we may be said to employ the sort of preservative memory that Burge advocated for, although I follow up my vindication of preservative memory with a brief discussion of why my ruling should not be offensive to those whose sensibilities are more closely aligned with Christensen and Kornblith. I conclude the second chapter by explaining how memory trace theory supports the Single Fragment Model as an abstract representation of idealized memorial mechanisms, despite the obvious

superficial differences between the ways in which they conceptualize our memorial processes. While memory depends on a reconstructive process, it is able to reconnect us with mental state contents which are, in terms of their contents and justifications, past mental states.

My final chapter begins by introducing the Preservation View of Memory, which is entailed by the Memory Justification Principle that I propose in chapter two, and which Jennifer Lackey believes to be false. The basic idea behind the Preservation View of Memory is that the epistemic status of the original counterpart of any memory-belief will always be greater than or equal to the epistemic status of the memory-belief itself. Lackey's argument against the Preservation View of Memory depends on the possibility that normative defeaters can effect changes in the epistemic statuses of beliefs (including memory beliefs) without their believing agents acquiring any further evidence relevant to those beliefs. The notion that as epistemic agents we can be subject to normative defeaters depends on the possibility that we have epistemic obligations while behaving as epistemic investigators. After unpacking these notions by way of Hilary Kornblith's discussion of epistemically responsible agency, I find that the existence of normative defeaters is not immediately objectionable, and determine that Lackey's case against the Preservation View of Memory (and, by extension, my Memory Justification Principle) succeeds. I reformulate the Memory Justification Principle as the *Internalized* Memory Justification Principle by simply adding that the Internal Memory Justification Principle only speaks to internal justificatory influences, such that the Internalized Memory Justification Principle preserves the basic intuitions behind the Memory Justification Principle while being compatible with the results of Lackey's normative defeater cases. Furthermore, I note that Lackey's normative defeater cases rely on an intuition that is roughly the Internalized Memory Justification Principle. I conclude by discussing two rather curious consequences of the Internalized Memory Justification Principle, which I ultimately argue are not as counterintuitive as they might seem at first, since these consequences simply demonstrate that the Internalized Memory Justification Principle supports a thoroughgoing virtue epistemology, in which justified memory-beliefs are the long-term rewards of virtuous acts of belief formation, and unjustified memory-beliefs are the long-term consequences of vicious acts of belief formation. We thus see that memory, as an internal mechanism, preserves internal sources of justification for the beliefs which it reconnects us with.

While these three chapters present only a brief survey of the philosophical issues pertaining to memory, I believe that they are sufficient to reveal that considerations of memory are indispensable for our epistemological theorizing.

Chapter 1

Fragmented Belief Structures and the Single Fragment Model

1.1 Fragmented Belief Structures

One task facing the formal epistemologist is the matter of addressing the problems that might arise when agents unwittingly harbor logically inconsistent beliefs within their belief structures. If belief structures are taken to be sets of beliefs closed under classical logical implication, then an agent who unwittingly believes both p and $\neg p$ is also committed to the belief that $p \wedge \neg p$, and is therefore taken to be committed to all expressible propositions, *ex falso quodlibet*. This is problematic, for even the most rational among us have likely believed contradictory propositions at certain points in our lives, but to think that we are all committed to believing *everything* is preposterous. The challenge facing the formal epistemologist is to develop a logical analysis of agents' doxastic structures that is able to escape the issues arising from agents' holding contradictory beliefs, while also appealing to our intuitions about how actual agents' beliefs are structured so as to avoid concerns of merely laying an *ad hoc* patch over the issue.

David Lewis's "Logic for Equivocators"[1] explores one avenue for dealing with the problems raised by classical models of belief structures. Citing a suggestion by J. Michael Dunn[2], Lewis establishes four desiderata for "truth and falsity according to some corpus of information": (1.) All explicitly affirmed propositions are true according to the corpus that affirms them, (2.) The explicit affirmations of a corpus are not its only truths, but "to some extent" truth according to the corpus is closed under logical implication, (3.) Contradictions or inconsistencies in the corpus do not necessitate that the corpus contains all expressible propositions, and (4.) Falsity of a sentence according to a corpus is defined as its negation being true according to the corpus (435). Lewis departs from Dunn, however, by integrating the notion of a corpus of information into our models of actual agents' belief structures with the

[1] Lewis, David. 1982. "Logic for Equivocators." *Nous* 16: 431-441.

[2] Attributed to: J. Michael Dunn. 1976. "Intuitive Semantics for First-Degree Entailments and 'Coupled Trees'." *Philosophical Studies* 29: 149-168.

notion of belief *fragmentation* (436). Speaking to the structure of an agent's corpus of beliefs, Lewis writes:

> The corpus is fragmented. Something about the way it is stored, or something about the way it is used, keeps it from appearing all at once. It appears now as one consistent corpus, now as another. The disagreements between the fragments that appear are the inconsistencies of the corpus taken as a whole. We avoid trouble with such inconsistencies . . . by not reasoning from mixtures of fragments. *Something is true according to the corpus if and only if it is true according to some one fragment thereof* . . . What follows from two or more premises drawn from disagreeing fragments may be true according to no fragment, hence not true according to the corpus (436, my emphasis).

With the notion of a fragmented belief corpus, Lewis allows for an agent's complete set of beliefs to be considered as a collection of smaller "fragmentary" belief sets. In this way each of an agent's belief fragments may be thought of as closed under unrestricted classical implication, while the complete structure (*i.e.*, the belief corpus) of which each fragment is a part is not closed under implication and may thus store inconsistent beliefs (437).

To see how Lewis' fragmentation theory approaches solving the problem of inconsistent belief structures, we may consider the following situation. Suppose that I believe that *Apocalypse Now* is the best movie I have seen, that Francis Ford Coppola directed *Apocalypse Now*, and that the best piece of Francis Ford Coppola's work that I have seen is *The Godfather*. Let us assume that I also have a tacit belief that *Apocalypse Now* and *The Godfather* are distinct films—that is, I have never explicitly affirmed "*Apocalypse Now* and *The Godfather* are not the same movie," but my conceptualizations of the two movies is incompatible with their being one movie. The conjunction of my aforementioned beliefs is incompatible, then, since I believe that *Apocalypse Now* is superior to all other movies and directed by Coppola, that of all the movies Coppola has directed *The Godfather* is superior, and I would reject the possibility that *Apocalypse Now* and *The Godfather* are identical. If I were alerted to the fact that I on one occasion affirmed that Coppola's best work is *The Godfather* but that I have also remarked that no movie is better done than *Apocalypse Now*, I would notice the need to change my beliefs; I would realize that because I know that *Apocalypse Now* was directed by Coppola, my belief system ought to reflect that Coppola's best work is not *The Godfather* (or that *Apocalypse Now* is not the best movie that I have seen), and thus adjust my beliefs to eliminate the inconsistency. My mistake is simply that I have never realized that Coppola directed both *The Godfather* and *Apocalypse Now*, and have thus never detected the conflict in my beliefs.

The utility of Lewis' fragmentation theory is that it protects one's corpus of beliefs from the threat that some minor inconsistency in one's complete belief set would necessarily commit one to all expressible propositions, *ex falso quodlibet*. So long as whatever contradictory beliefs that I hold are not members of the same fragment, then, I am not committed to all expressible propositions. Suppose that my beliefs that *Apocalypse Now* is the best movie I have seen and Coppola directed *Apocalypse*

Now are in a separate fragment from my belief that *The Godfather* is Francis Ford Coppola's best work. These separate fragments are themselves closed under logical implication, such that in one fragment I hold that Coppola's best movie is *Apocalypse Now*—supposing that this follows logically from "*Apocalypse Now* is the best movie I have seen" and "Coppola directed *Apocalypse Now*"—whereas in the other fragment I hold that Coppola's best movie is *The Godfather*. My complete belief corpus, which contains both of these fragments, is not closed under logical implication, and I am thus able to believe: (i.) "Coppola's best movie is *Apocalypse Now*" and (ii.) "Coppola's best movie is *The Godfather*," without believing the contradictory claim (iii.) "Coppola's best movie is *Apocalypse Now* and Coppola's best movie is *The Godfather*." Propositions (i.) and (ii.) are true within fragments of my corpus, and thus true in my corpus, whereas (iii.) is not true of any fragment within the corpus, and is thus by definition not true in the corpus. So long as this is the case I am not committed to believe everything, even though my corpus of beliefs contains beliefs that are logically inconsistent when considered together.

Andy Egan's paper "Seeing and Believing: Perception, Belief Formation and the Divided Mind" argues, in part, for a similar notion of fragmented belief modeling.[3] The problem that Egan wishes to address with fragmentation is in the same vein as that which inspired Lewis: "On many of the idealized models of human cognition and behavior in use by philosophers, agents are represented as having a single corpus of beliefs which (a) is consistent and deductively closed, and (b) guides all of their (rational, deliberate, intentional) actions all of the time" (48). Egan, like Lewis, finds this picture unrealistic and untenable, and pulls from Lewis the idea that "we have a number of distinct, compartmentalized systems of belief, different ones of which drive different aspects of our behavior in different contexts" (48). In order to fully develop this notion, Egan characterizes "the behavior-guiding role of belief and desire" in this model as the dispositions of agents "to act in ways that would satisfy their currently-active desires if their currently-active beliefs were true," which enables Egan to define belief in the following way: "[For any agent x and proposition P] x believes P iff x is in some state that represents that P and is disposed, *when active in behavior-guidance*, to cause x to act in ways that would be successful if P" (52, my emphasis). The utility of this definition of belief is that it is allows for an agent to believe a proposition while not requiring that she behave as though that proposition were true—specifically, it allows that the explicit beliefs within an agent's belief corpus that are not part of her active fragment nevertheless qualify as her beliefs. We can thus believe all the propositions that constitute our belief corpuses, while only ever having a comparatively small subset of these beliefs governing our thoughts and behavior at any given time.

Arguing for the plausibility of his theory of fragmented belief structures, Egan remarks, "What we see in actual believers looks more like (and is most charitably interpreted as) a pattern of more-or-less rational updating of fragments, in which not every update effects every fragment" (55), whereby fragments become updated based

[3]Egan, Andy. 2008. "Seeing and Believing: Perception, Belief Formation and the Divided Mind." *Philosophical Studies* 140: 47-63.

on intrinsically specified "belief-updating mechanisms whose deliverances they're sensitive to" (52). Much of "Seeing and Believing" is focused on the argument that in addition to the fact that actual agents have fairly limited cognitive resources—which a fragmented concept of belief captures better than the traditional single corpus model—fragmentation models promise to capture more of the idiosyncrasies in our cognitive behavior, such as how we frequently express varying credences towards certain of our beliefs depending on our particular contexts (55). Furthermore, while Egan's picture of the revision and updating process is somewhat gestural, the basic idea is clear: our belief fragments have certain sensitivities that make them vulnerable to change based on the influence of new beliefs or the contents of our perceptions (58). There remains much more to be said on the matter of belief updating and revision, but what is important is that Egan's argument shows that there do not appear to be any strong *prima facie* reasons to deny the plausibility of belief fragmentation as a significant feature of our actual cognitive structure. Egan has thus expanded upon the basic idea of a fragmented belief corpus (as developed by Lewis) by providing a definition of belief that is compatible with fragmentation, as well as remarking on how examining particular aspects of an agent's cognitive context will be essential to understanding the activation and revision processes that operate on his or her belief fragments. The upshot of Egan's contribution for the theory of fragmented belief structures is that it provides an intuitive argument that the fragmented belief structures approach to modeling cognition is a natural way to model the cognitive activities of actual agents, showing that this solution to the problem of agents possessing inconsistent beliefs is not *ad hoc*.

There are, nonetheless, certain aspects of Egan's characterization of fragmented belief structures that require further examination. The notion that there are "more-or-less rational" mechanisms behind the activation and updating of particular belief fragments could, in particular, benefit from more development, so that we might have a sound account for why specific fragments would be active and subject to change within certain contexts and not others. Furthermore, while Egan shows that a model which characterizes agents as having a fragmented belief structure is more natural than the traditional single belief set model, an argument can be made that the ontological robustness of the fragments within Egan's model does more to harm his project of naturalizing epistemic modeling than to help it, since it is not immediately apparent that our beliefs are rigidly compartmentalized. There are benefits to modeling our belief structures as fragmented, but if these models do not agree with our general understanding of the mind then we cannot appeal to fragmentation in order to escape the problems that arise from inconsistencies within agents' belief sets. With these considerations in mind, I intend to work towards establishing a more fully developed, psychologically plausible model of fragmented belief structures.

1.2 Relevance Sensitivity and Querying

Samir Chopra and Rohit Parikh develop a thoroughgoing formal model of fragmented belief structures in their paper "Relevance Sensitive Belief Structures."[4] This model is focused around Chopra and Parikh's notion of B-structures, where a B-structure is "a collection of sets of beliefs, each one of which is individually closed and consistent. Their union *may* be inconsistent, however, and querying ... can expose this inconsistency under the right conditions" (9, their emphasis). B-structures are thus analogous to the belief corpuses that Lewis advocates for, and each set of beliefs within a B-structure is like a belief fragment. Taking finite propositional languages, which are sets of propositional atoms (3), and theories, which are sets of formulae made from propositional languages (3), Chopra and Parikh propose the following definition of B-structures:

> A belief structure B on [a propositional language] L is a set $\{(L_1,T_1), \ldots, (L_n,T_n)\}$ such that $L = \bigcup L_i : i \leq n$, and each T_i is a consistent, finitely axiomatizable theory in L_i. The T_i are $Cn(\Gamma_i)$[5] where the Γ_i are the explicit beliefs of the agent in language L_i. If $n = 1$, then B will be just a theory (9).

In less formal terms, a B-structure is a set of ordered pairs of propositional languages and the logically closed theories (*i.e.*, beliefs) which take the elements of those propositional languages as atomic. Considering B-structures as analogous to Lewis's belief corpuses, the ordered pairs in B-structures function analogously to the individual belief fragments that Egan describes.

Chopra and Parikh refer to this as a partial language splitting model, since a B-structure with $n > 1$ will have n theories that are each specified to a minimal[6] propositional sublanguage (hence language splitting), and the distinct sublanguages in the B-structure are allowed to share elements, such that the complete propositional language is not necessarily partitioned into mutually disjoint sublanguages (hence partial) (8). The B-structures model is inconsistency-tolerant in virtue of allowing for the overlap of sublanguages within a B-structure (9), since the overlap of sublanguages is required for the production of inconsistent theories—were this not the case, the theories could not be in conflict over the same subject matters,[7] and thus could not be inconsistent. We can see that, thus far, Chopra and Parikh have established an inconsistency-tolerant theory of belief structuring that explains the organization of individual belief fragments by appeal to the notion of language splitting, whereby the

[4] Chopra, Samir and Rohit Parikh. 2000. "Relevance Sensitive Belief Structures." 1 - 25. www.sci.brooklyn.cuny.edu/ schopra/maidone.ps. Internet.

[5] Cn is the function that takes a set of formulae (*i.e.*, a theory) as its argument and returns a set of formulae equivalent to the logical closure of the input set (3).

[6] Minimal meaning that the language is the smallest set of propositional atoms required for the expression of the respective theory.

[7] The subject matter of a proposition is, for Chopra and Parikh, the "language of the formula used to express the proposition" (8), and is thus a way to speak to arbitrary subsets of propositional languages.

fragments are defined relative to specific subject matters (*i.e.*, sets of propositional atoms) and thus group theories according to what they are fundamentally about.

Emphasizing the psychological plausibility of their model, Chopra and Parikh point out that the groupings of formulae that establish the theories of a B-structure will be subjectively determined, writing, "We place beliefs in subsets according to how relevant[8] we take them to be to one another . . . We might inspect an agent's B-structure and find that, objectively, a different splitting could have been carried about" (10). For example, consider the following four formulae: "Snow is white and snow is cold," "Snow is white and printer paper is white," "Printer paper is white and ink is black," and "Snow is white and ink is black." Each of these formulae are expressible in the language $L = \{$snow is white, snow is cold, printer paper is white, ink is black$\}$. It is no more correct for an agent to partition her language into sub-language $L_1 = \{$snow is white, snow is cold, printer paper is white$\}$ and sub-language $L_2 = \{$printer paper is white, snow is white, ink is black$\}$, such that $T_1 =$ "Snow is white and snow is cold. Snow is white and printer paper is white." and $T_2 =$ "Printer paper is white and ink is black. Snow is white and ink is black." than to partition her language into $L_1 = \{$snow is white, snow is cold, ink is black$\}$ and $L_2 = \{$snow is white, printer paper is white, ink is black$\}$, such that $T_1 =$ "Snow is white and snow is cold. Snow is white and ink is black." and $T_2 =$ "Snow is white and printer paper is white. Printer paper is white and ink is black." Moreover, a different partitioning could be done so that the agent had a sublanguage for each formula, or the agent might not have any partitions with respect to language L. Within Chopra and Parikh's partial language splitting model, the particular structuring of an agent's sublanguages are simply determined by how her psyche decides to organize her theories. In terms of belief corpuses and fragments, the B-structures model allows for an agent's (possibly inconsistent) corpus of explicit beliefs to be broken into individually consistent fragments, which are organized around her or his subjective assessments about which beliefs are most relevant to one another.

Recall the inconsistency within my beliefs that "*Apocalypse Now* is the best movie I have seen," "*The Godfather* is the best movie I have seen that was directed by Coppola," and "*Apocalypse Now* was directed by Coppola." The B-structures model can explain this situation by claiming that I must not find the subject matters of these beliefs sufficiently relevant to group them into a single theory. This is an intuitive explanation of how such a situation can arise. Perhaps I have grouped all my beliefs about the purely artistic features of *Apocalypse Now* and my subsequent belief that *Apocalypse Now* is the best movie I have seen into one theory, while I have a separate theory organized around my more strictly Coppola-relevant beliefs, which includes my beliefs that *The Godfather* is Coppola's best work and that Coppola directed *Apocalypse Now*. Because my particular psychological features do not closely associate the artistic features of *Apocalypse Now* with Coppola, I have failed to recognize my

[8]Earlier in their paper, Chopra and Parikh define the relevance of propositional formula as follows: "Formulas α, β are *relevant* to each other, $R(\alpha, \beta)$ iff their smallest languages L_α, L_β share propositional atoms, i.e., iff $L_\alpha \cap L_\beta \neq \emptyset$" (5). This is an objective measurement of whether two formulae are relevant to one another, and is not taken to weigh in on the subjective assessment of how they ought to be grouped in an agent's B-structure.

inconsistent belief; I would have to participate in a query[9] that causes both of these theories to activate before I could recognize the inconsistency and revise my beliefs accordingly. Chopra and Parikh's emphasis on the notion of subjective relevance assessments provides an intuitive account of how fragments become organized—namely, in accordance with what subject matters (*i.e.*, propositional atoms) the agent prefers to group his or her beliefs—that also conforms to the types of stories one might offer to explain how an agent could end up with inconsistent beliefs. Relevance assessments are thus capable of playing a crucial role in developing an intuitive model of fragmented belief structures, given their ability to explain how belief fragments are organized as well as how queries are able to cause the activation of specific fragments.

1.3 *B*-structure Revision and Psychological Plausibility Concerns

"There are two strategies available to an agent for handling belief revision on belief structures . . . In each case, the revision procedure will be sensitive to the smallest language that the new epistemic input is expressed in," write Chopra and Parikh, introducing their section on belief revision (16). By having two revision strategies that append theories based on the relevance of new information to an agent's preferred subject matter organizations, Chopra and Parikh clearly aim to appeal to the same sorts of intuitions that make their relevance sensitivity account of fragment organization and querying psychologically plausible. The first option that Chopra and Parikh introduce, Option A, is to "Revise all sub theories relevant to new information *without merger*" (16, my emphasis). The intuition behind this strategy is that it allows us to handle cases where incoming information is relevant to multiple theories that may not be relevant to one another. The example that Chopra and Parikh use is that "If I learn that both Beijing and London had cold winters, I am not likely to merge all my other beliefs about the two cities" (16). The second revision option is Option B, which will account for "Revision which involves the merger of theories" (16). Chopra and Parikh wish for Option B to handle cases where "an agent *repeatedly* receives some information which overlaps two sub-languages. The agent may decide that the division is artificial and should be abandoned" (16). An art theory student, for example, might have a set of beliefs about the concept of the image, and a different set of beliefs about the concept of the representation, but after attending a series of lectures form the meta-belief that his beliefs about the image and his beliefs about the representation speak to the same thing, and on this basis wish to have his future theorizing reflect this realization. Option B is useful because it allows for agents' belief structures to react to these sorts of realizations.[10]

Chopra and Parikh's two options for revision are formally satisfying insofar as they

[9] A query can be thought of as the presentation of a formula to an agent (either in the course of his own cognitive processes or by an external source), the linguistic features of which stimulate the activation of a particular theory or theories (12).

[10] Chopra and Parikh give much more detailed accounts of these revision methods in their paper (17-9), but I will leave such details alone for the sake of my present inquiry.

can be used to analyze all cases of belief revision, but the introduction of two revision mechanisms seems vulnerable to criticisms of being psychologically implausible or *ad hoc*. There is little evidence supporting the theory that the types of belief revision activities that occur when an agent learns that London and Beijing both had cold winters and when an agent realizes that the image and the representation are the same thing are actually different kinds of activities. If there were strong evidence to this effect, it would seem that the traditional idealization of a single belief base would not have held as much traction as it did, and we would have done more to incorporate the distinction between A- and B-type revisions into the formulations of our traditional models of belief structures. Looking closely at the A- and B-type revision methods, the intuition that appears to be guiding Chopra and Parikh in distinguishing the two methods of revision is that the acquisition of new beliefs can impact how one might be expected to (or want to) respond to situations in the future. When I learn that Beijing and London had cold winters, my collection of Beijing-relevant beliefs and my collection of London-relevant beliefs are said to each gain one more proposition because it is unlikely that on the basis of this one new belief I would be expected to always think of Beijing whenever I thought of London, and vice-versa. This is to say that I would not expect my belief that Beijing and London had cold winters to generate relevance connections between all of the formulae in my Beijing-relevant fragment and all of the formulae in my London-relevant fragment. Similarly, when I learn that the image and the representation are the same thing, my belief corpus (B-structure) gains a proposition that explicitly reveals the relevance of my "image" and "representation" fragments to one another, and thus it becomes desirable for me to always activate my "image" beliefs simultaneously with my "representation" beliefs. The effect of dividing belief revision into the distinct A- and B-types of revision appears necessary for the B-structures model's adequate characterization of how actual agents revise their beliefs, but given that the division of the mental processes underlying actual belief acquisition into distinct A- and B-type mechanisms does not appear psychologically plausible,[11] the B-structures model's account of belief revision essentially reveals its own artificiality.

The psychological implausibility of the B-structures model's approach to belief revision raises deeper issues for the project of modeling agent's belief structures through fragmentation. We saw that the rather sophisticated B-structures model required positing multiple revision processes in order to capture our intuitions about the different ways in which belief acquisition actually impacts agents' doxastic structures. If we accept the general framework of fragmentation used by Egan and Chopra and Parikh, then we are thus stuck between two undesirable alternatives: we can either allow our models multiple belief revision processes (sacrificing psychological plausibility with respect to how the mind treats incoming information) or else build our models around a single belief revision process (sacrificing plausibility with respect to our experiences of belief acquisition). I am unwilling to accept either of these

[11] At the very least, claiming that actual belief revision is always performed through one of two distinct mental processes on the basis of the perceived utility of the inserted belief for future reasoning is suspiciously convenient for the B-structures model.

consequences, and thus maintain that the only way to proceed with developing a psychologically plausible model of fragmented belief structures is by rejecting certain features of the framework laid out by Chopra and Parikh. In order to begin reconfiguring our general approach to modeling belief using fragmentation, we may benefit from noting that the reason why Chopra and Parikh's B-structures model requires two revision mechanisms is ultimately due to its rigid organization of agents' B-structures into ordered pairs of sub-languages and theories. Were it not for the fact that the B-structures model gives so much ontological weight to the fragments that it contains there would be no need to have different methods of updating. It thus appears that the best approach to enhancing the psychological plausibility of the fragmented belief structure notion will involve relaxing the degree to which an agent's beliefs are taken to form coherent fragments. The challenge before us is to develop a fragmentation-based model of belief that enjoys all of the formal benefits of the B-structures model, but which approaches belief fragmentation in such a way that does not compromise psychological plausibility in the presentation of its essential mechanisms—in particular, its characterization of belief revision.

1.4 The Single Fragment Model

The need for multiple types of revision mechanism is a product of having a robustly fragmented belief structure. Both Egan's and the B-structures model's approaches to fragmentation are robust in the sense that an agent's entire belief structure is taken as being broken into collections of subject-specific theories *at all times*. I wish to propose an alternative model, which I will refer to as the Single Fragment Model (hereafter SFM). The essential idea behind the SFM is that it allows agents to have only two identifiable collections of beliefs: a set containing all of and only the agent's explicit beliefs, and a variable subset of those beliefs which is taken to be closed under classical logical implication. Because the primary benefit of belief fragmentation is that it allows an agent to have epistemic access to the complete logical implications of certain of his beliefs at any time while tolerating inconsistencies within his or her total set of beliefs, it will first be necessary to organize the SFM around this capacity to tolerate inconsistencies within agents' beliefs.

Let an agent's *belief corpus* be the set of all of his or her explicit beliefs,[12] and let an agent's *active fragment* be the belief set constituted by the logical closure of the subset of the agent's explicit beliefs which are involved in the agent's current cognitive activities (theorizing, decision-making, etc.). The propositions in the agent's active fragment that are not elements of his or her belief corpus are the agent's *implicit beliefs* (or commitments). Assuming that the propositions contained in an agent's belief corpus must be constructed from the atoms of some finite propositional language, the elements of an agent's active fragment will always be expressible by a sub-language of the agent's complete propositional language. This allows for the same sort of querying that Chopra and Parikh's model used, only instead of activating entire theories at a time based on the language of the query, individual beliefs within the agent's corpus

[12]Maintaining the fragmentation-oriented definition of belief provided by Egan on (52).

are activated (based on their relevance[13] to the query), such that the logical closure of the activated beliefs constitutes the active fragment. The SFM is thus able to contain all of the explicit beliefs of an agent within the agent's belief corpus while theoretically providing him epistemic access to the complete logical consequences of any subset of those beliefs[14] *via* his active fragment. In this way, the SFM is able to tolerate inconsistencies in the belief corpuses of agents in the same way that Chopra and Parikh's B-structures model does, and thus achieves the basic desiderata for a fragmentation model of agents' belief structures that Lewis presented in "Logic for Equivocators."

The notion of relevance used by the B-structures model will require a significant amount of naturalization in order to accommodate the myriad triggers that can cause certain beliefs to activate for actual agents. Recall that the most basic elements of the languages considered by the B-structures model are propositional atoms. For the sake of developing a functional formal model, taking atomic propositions as the smallest components of a language is certainly reasonable, but in cases involving actual agents the possible bases for connections between beliefs cannot be arbitrarily restricted in this way. Not only are syntactic components much more specific than proposition letters (*e.g.*, predicates, designators) capable of establishing connections between the beliefs of actual agents, but some connections cannot even be properly understood by examination of the relevant beliefs' semantic features. Suppose, for instance, that while I am backpacking in the wilderness I spill a container of instant cappuccino powder all over the inside of my backpack, and for the duration of my expedition all of my equipment smells like instant cappuccino. After a couple of days of constant exposure to the cappuccino smell, I form the belief "Always double-bag powdered goods for long backpacking trips." Years later, I might be buying groceries with a housemate and experience a smell that is similar to the smell of the instant cappuccino, and on the basis of my present olfactory experience recall how I smelled like instant cappuccino for the better part of a backpacking trip. Further, I may comment to my housemate, "The next time you go backpacking be sure to double-bag your powdered stuff," as a lead-in to my story about smelling like instant cappuccino. In order for the SFM's account of relevance querying to be sufficiently naturalized for modeling actual agents, it must be able to explain how it is that something as slippery as a component content of one's present perceptual experience (*e.g.*, the smell of some instant cappuccino) can cause the activation of beliefs from one's corpus. This reveals that in order to establish an adequate account of relevance, we will first need to examine what the SFM should take as the basic constituents of beliefs.

According to Egan's definition of belief, a proposition p is believed by an agent

[13]Relevance is taken to be involved in all querying within the SFM, but not the only explanatory factor for what beliefs become activated. In the following section I will present the notion of centrality as a more fine-tuned approach to the matter of belief activation.

[14]While the SFM allows for the possibility that an agent's active fragment contains his entire belief set and its logical consequences, actual agents (certainly all human agents) may be assumed to have certain psychological limitations that would prevent them from activating too large of a swath of their belief corpus at any given time.

if (1.) that agent is in a state that represents that p, and (2.) p is such that when active it would dispose the agent to act in a way that would be successful if p were true (Egan, 52). The first task for the SFM, then, is to explain how contents can be introduced into agents' belief corpuses which qualify as beliefs and which can contain constituents that have semantic content, as well as constituents that have other kinds of contents (*e.g.*, sense-perceptual contents).

Recall that in order for a belief to be stored in an agent's corpus, it is only necessary that the agent explicitly affirm that belief. In order for experiential contents to be stored in an agent's belief corpus, it will thus be necessary for agents to have some means for affirming portions of their complete mental state contents at any given time. One way that agents can plausibly do this is through the use of demonstrative concepts. Suppose that while I am backpacking, reeking of instant cappuccino, I have the thought "Always double-bag powdered goods for long backpacking trips," and follow that thought up by saying to myself, "*This* is not what I planned." With my second utterance ("This is not what I planned") I use the demonstrative 'this' to refer to my mental state at that time—a mental state that includes the cappuccino smell as a partial content. By affirming a proposition containing a demonstrative concept that refers to some portion of the content of one's present mental state, one is able to "package" a portion of that mental state's content as part of an affirmed belief, and thus admit parts of one's mental states (possibly including particular sense-perceptual contents) into one's belief corpus. In the backpacking case, my use of the demonstrative "this" may therefore be thought of as standing in for a proposition along the lines of "I smell like $s \wedge$ I have realized 'Always double-bag powdered goods for long backpacking trips' is true $\wedge$ ...," where s stands for the smell of the instant cappuccino and the length of the proposition simply depends on how much of my present mental state's content "this" captures. The proposition that I have affirmed ("This is not what I planned") is represented by my present mental state, and can be assumed to be such that it would dispose me to perform successfully if assumed true, so what I have affirmed can qualify as a belief under Egan's definition.[15] Because the smell of the cappuccino functions in this case as a sub-propositional constituent of a belief that I have affirmed, it may be stored along with part of that belief as an element of my belief corpus. A similar method can be applied to any kind of mental content, and thus the SFM is able to account for the storage of the contents of any kind of sub-propositional propositional constituents.

Given these considerations, the best route to naturalizing relevance connections for the SFM appears to be through an appeal to relevant beliefs' sharing sub-propositional propositional constituents. We can maintain the structure of Chopra and Parikh's approach to defining relevance (Chopra and Parikh, 5), while abstracting away from concrete syntactical elements and into the language of content through use the following definition of the relevance of two propositional formulae: Propositional formulae a and b are relevant to one another iff they share sufficiently similar sub-

[15]While I incorporate "this" into an explicitly affirmed natural language proposition, this is merely for ease of illustration. Presumably, the majority of cases in which we store these "packages" of sense-perceptual contents do not actually involve their being embedded in linguistic propositions, and are rather affirmed in a more gestural way.

propositional propositional constituents, where sufficient similarity is subjectively determined (*i.e.*, determined by one's particular psychological features) but roughly approximates equivalence.[16] Like the B-structures model's account of relevance, the notion of relevance used by the SFM is also a binary symmetric relationship between propositions within a belief structure—two beliefs in a belief corpus are either relevant to one another or they are not, and if an agent's belief a is relevant to her belief b, then b is necessarily relevant to a for that agent. The difference between the two accounts of relevance is simply what they take the essential constituents of belief to be. As a rigidly structured formal model, the B-structures model reasonably considers propositional atoms as the smallest constituents shared between theories, whereas the SFM, as a naturalized model of belief fragmentation, takes nonconceptual content as what can establish relevance between an agent's beliefs, and thus provides the basis for objective relevance assessments.[17]

Besides matters specific to the two models' notions of relevance assessment, the process of querying within the SFM remains almost identical to that used in the B-structures model, insofar as the language of a query is taken to cause a certain relevant subset of the agent's explicit beliefs to activate. Suppose, for example, that I am asked "Is it true that horses have four legs?" In the B-structures model, I would activate the theory that I have paired with a language containing the proposition "Horses have four legs," and respond to the query on the basis of whether my relevant theory affirms, affirms the negation of, or takes no stance regarding the proposition "Horses have four legs." In the SFM, I might begin by activating my beliefs that are relevant

[16] In the case where I recall that the smell I experience in the supermarket is similar to how I smelled when I was backpacking, the two smells do not have to have precisely the same characters in order for my memory of the backpacking experience to be activated. It is easy to imagine, for instance, that the instant cappuccino in the store is vanilla flavored, whereas the cappuccino powder I took backpacking was hazelnut flavored, but my relevance-querying mechanism nevertheless activates my backpacking belief about double-bagging. It is this sort of sameness of content characteristics that I intend to capture with the notion of sufficient equivalence-approximating subjectively-determined similarity.

[17] One side note regarding the SFM's basing of relevance on the content of beliefs is that it requires some minor qualification regarding the matter of the language of querying. Because relevance determines the subject matter of a query, the literal subject matters of the SFM are the sub-propositional constituents of beliefs, which in some cases may evade strict linguistic confinement. That said, it seems plausible that the languages of most agents are such that, for the sake of describing the querying process, the sharing of certain key concepts by beliefs can be taken as a standard indicator of relevance. For example, though the propositions "Horses have four legs" and "Horses are mammals" are most immediately understood as relevant to one another in virtue of sharing the concept that is typically associated with the English word 'horse,' it is possible that the actual content of the beliefs has nothing to do with the sound of the word 'horse,' and thus in a certain sense 'horse' is not the genuine constituent of the query, but whatever content corresponds to the word 'horse' for a particular agent. Nevertheless I assume that the content that we might use the term 'horse' to capture will typically line up sufficiently well with the use of the term 'horse' that for the sake of describing instances of relevance querying it is safe to use phrasing along the lines of "Agent A performed a horse-belief query" to describe what happens when A is asked if it is true that horses are four-legged mammals. Throughout the rest of my introduction of the SFM I will thus be using a phrases such as "a 'horse'-based query" or "beliefs relevant to an agent's notion of 'horse'" in a qualified way—namely, as a shorthand for the sort of gesturing that would be required to speak to the content that actually determines the relevance of two propositions.

to the primary contents of the query, most likely 'horse' and 'legs,' and stop once I have activated a belief that affirms or affirms the negation of "Horses have four legs," or once I somehow become confident that I have no beliefs regarding whether horses are quadrupeds, and abandon the query. The SFM's approach to querying is only more natural than the B-structures model's approach in virtue of the SFM's more natural account of what determines the relevance of beliefs.

As it has been developed thus far, the SFM has been shown able to account for belief activation in a way that is much more similar to actual human cognition than the B-structures model, primarily by using a much more abstract notion of relevance and by doing away the B-structures model's commitment to ontologically rigid language-determined theories. There is, however, more work to be done to make the SFM an optimally naturalized model of fragmented belief structures. Consider, for example, an agent who is politically active, and exerts a significant amount of cognitive energy thinking about propositions involving the terms 'government' and 'should.' If such an agent were to be asked what she thinks the government should be doing, it is much more plausible that she would first activate the belief that she finds at that time to be most important to her ideas about what the government should be doing, then the next most relevant, and on down the line, rather than instantly have her complete list of beliefs about what the government should be doing instantly cognitively available. As it has been developed thus far, however, the SFM does not necessarily give this result. Insofar as querying uses only the notion of relevance as developed by Chopra and Parikh, the political agent would only activate beliefs based on their sharing propositional elements with the language of the query, and thus the order of the beliefs activated by the agent would not necessarily reflect which she might deem most important. In the following section, I will develop a notion of belief centrality that will be crucial for providing an adequate single-mechanism approach to belief revision within the SFM, but also gives the desired result that beliefs that seem most important to an agent will be given priority during a relevant query, which will further demonstrate the SFM's ability to characterize cognitive processes in a psychologically realistic way.

1.5 Centrality and Belief Revision in the Single Fragment Model

Issues with the psychological plausibility of the B-structures model's account of belief revision were the inspiration for seeking out an alternative model of cognition, and we may now examine whether the SFM achieves greater plausibility with respect to its account of belief revision. We defined the belief corpus of an agent as the set of all of his explicit beliefs, so whenever an agent explicitly affirms some proposition that proposition is simply added into his belief corpus. When I form the belief that both Beijing and London had cold winters, my belief corpus goes from not containing the proposition "Beijing and London had cold winters" to containing it. When I form the belief that 'image' and 'representation' are synonymous, a corresponding

proposition is simply added to my belief corpus. The issue that the B-structures model ran into with even the simple amendment component of belief revision is that its robust division of the belief corpus into subject-specific theories required that the revision mechanism somehow make sense of what sort of amendment was being made before determining what process to use. The SFM bypasses this problem by basing its account of revision on the brute insertion of beliefs into an agent's belief corpus, and thus simplifies the process of acquiring beliefs in a way that is at least initially more psychologically plausible than the B-structures model's account. There is, however, much more to belief revision than belief acquisition, and the new challenge for the SFM is to explain how *robust* belief revision—that is, revision other than belief acquisition—can occur in a psychologically plausible way within the present setup of the SFM.

Consider, for example, an agent who believes that his living room rug is purple. One day, while entertaining company in his living room, one of the agent's friends remarks, "That is a nice fuchsia rug you have." The agent has never ascribed a more precise color term to the rug than 'purple,' but realizes that 'fuchsia' is a much better word for the rug's color, and for this reason wishes to revise his belief structure such that whenever he is queried regarding his rug's color he will respond "I have a fuchsia rug," rather than "I have a purple rug." In order to handle these kinds of revisions, the SFM must appeal to more than relevance in accounting for the querying process used by agents. Recall that Chopra and Parikh define the relevance of propositional formulae as a boolean function on whether the formulae share propositional atoms; if they share propositional atoms, then they are relevant, otherwise they are not (5). In cases such as the agent with the fuchsia rug, the notion of relevance established by Chopra and Parikh cannot fully make sense of the sort of revision that is taking place—the agent's beliefs "My rug is fuchsia" and "My rug is purple" are equally relevant to the query "What color is your living room rug?" and thus without some assessment other than relevance, the agent essentially flips a coin between which of his two rug-color beliefs will activate first whenever he is queried in this way.

To resolve this issue, the SFM makes use of the notion of beliefs being more or less *central* to relevant queries. The centrality of a belief to a relevant query can be thought of as the likelihood that it will become active in the course of that query. Centrality is a highly subjective[18] matter, and may be influenced by a variety of factors, such as how recently the belief was activated, the frequency with which the belief is activated, strong emotional connections to the belief, and so on. Centrality is thus not determined by any features of the propositional content of the belief, but is determined by the agent's relationship with the belief. In the same way that the B-structures model allows the placement of beliefs into theories to be dependent on the particular psychological environment of the agent, the SFM does not impose any strict limitations on how the centrality of any belief to a relevant query is determined. Further, we can define a belief as having zero centrality to a query[19] if it does not

[18]The subjectivity of centrality does not imply that degrees of centrality are assigned consciously.

[19]Or at least near-zero centrality, so as to allow for odd cases where an agent erratically activates beliefs, such as in cases where an agent is under the influence of powerful drugs.

share any propositional atoms with the language of that query. Thus a belief can have non-negligible centrality with respect to a query iff it is relevant to that query. The utility of centrality is that it allows for certain of an agent's beliefs to tend to display priority over other of the agent's beliefs, even if they are equally relevant to certain querying languages.

Recall the politically active agent who has been queried regarding what she believes the government should be doing. We can expect that the agent's more central beliefs for a query involving key terms such as 'should' and 'government' will activate first. Supposing that the agent has recently spent a great amount of time thinking about economic inequality, and has been less focused on the topic of interrogation and torture, the SFM anticipates that the agent's beliefs regarding economic issues will be more central to the present query than her beliefs regarding interrogation policy, and thus expect that her response to the query will begin with a discussion of how the government should work towards building a strong middle class, rather than a discussion of the immorality of waterboarding. In addition to providing a more psychologically plausible method of querying, the notion of centrality allows the SFM to account for why we might expect that certain of an agent's beliefs will be more disposed to activate during wide-scope queries than other of the agent's relevant beliefs.

With the notion of centrality in mind, we may now return to the topic of belief revision and the case involving the agent who wishes revise his belief structure such that he will tend to recall that his rug is fuchsia, rather than that it is purple. To bring about such a revision of how he will respond to rug color relevant queries, the agent may manipulate the centrality of his belief that the rug is fuchsia, perhaps by repeating to himself, "My rug is fuchsia," or by performing some other behavior along these lines, so as to make his belief that he has a fuchsia rug more central to queries regarding rug colors. We may note that these sorts of behaviors are, in fact, how many actual agents go about enhancing the likelihood that they will be able to reactivate certain beliefs at a later time. Additionally, because centrality only represents a tendency to respond to queries in certain ways, the SFM is able to account for situations where agents struggle to produce the correct or most desirable response to a query. It is conceivable that the agent with the fuchsia rug might occasionally respond to rug color queries with "I have a purple rug," despite having worked to reinforce the centrality of his belief that the rug's color is fuchsia. This simply represents a case where the agent's belief "I have a purple rug" beat the odds (so to speak) and activated before "I have a fuchsia rug" could. The notion of centrality is therefore capable of making the brutish approach to belief introduction that the SFM uses psychologically plausible.

Revisions where an agent outright rejects one of his beliefs can be accounted for in the same way—namely, the negation of the old belief is inserted into the agent's belief corpus, and the centrality of that belief (the negation) is strongly reinforced. Recall the case in which I am assumed to believe that *Apocalypse Now* is the best movie that I have seen, that *The Godfather* is the best movie that I have seen that was directed by Coppola, and that Coppola directed *Apocalypse Now*. Suppose I am discussing movies with a friend, and I remark, "*The Godfather* is definitely Coppola's best work,

but of all the movie's I've seen *Apocalypse Now* comes out on top." When my friend asks if I am aware that Coppola directed *Apocalypse Now*, I will realize my mistake in virtue of having my belief that Coppola directed *Apocalypse Now* activate, and subsequently correct myself, perhaps by saying, "Oh, right. I suppose *The Godfather* is Coppola's second-best work." The centrality of my beliefs that "*Apocalypse Now* was directed by Coppola" and "*The Godfather* is my second-favorite Coppola movie" are then reinforced, and I will thus be disposed to respond to inquiries regarding my movie preferences in a more coherent way in the future.[20] According to the SFM, then, all belief revision is simply a matter of the insertion of beliefs into an agent's belief corpus or the adjustment of the centrality of beliefs relative to their relevant querying languages.

One consequence of the SFM is that it implies that agents' belief structures can be densely populated by beliefs that they have rejected or adopted revised versions of. While this may seem counterintuitive at first glance, it is not unrealistic. Consider an agent who recently bought a new stereo system, and is listening to music with a friend. On the coffee table in the room with the stereo is a *Dark Side of the Moon* CD. When the album that the two are listening to ends, the agent's friend picks up the CD case, walks over to the stereo, and puts the disc in the tray. The friend asks the agent "Where should I leave this?" while holding up the CD case, to which the agent replies, "You can just leave it on top of the player." The friend leaves the CD case on top of the player, and neither interact with the CD or case again that day. The next day, the agent's roommate might call the agent and ask if he knows where their shared copy of *Dark Side of the Moon* is. It is plausible that the agent would respond, "I think it's on the coffee table," even though he both witnessed and verbally acknowledged that the CD case had been moved to the top of the stereo the previous day. In such a situation, when the agent's roommate informs the agent that the CD is not on the coffee table, the agent may well say, "Oh, that's right, Dylan left the CD in the stereo and the case should be on top." This is a familiar kind of experience, and shows that actual agents do not immediately overwrite or erase prior beliefs as their cognitive situations change.[21] What is important is that the SFM's account of belief revision is psychologically plausible, even though its consequence that agents' belief corpuses may contain a large population of rejected beliefs may seem initially counterintuitive.

1.6 Concluding Remarks

At the outset of this chapter, it was noted that traditional models of belief structures face a serious problem insofar as they commit agents who hold inconsistent beliefs to

[20]It is possible that the negativity of having experiences where one realizes that one's belief structure is incoherent might reduce the centrality of the incoherent beliefs. Just as certain activities might reinforce the centrality of certain beliefs to certain querying languages, it is plausible that other activities can weaken belief centrality.

[21]There might, perhaps, be some sort of "garbage collection" mechanism in the mind that does eliminate beliefs that meet certain criteria for uselessness, but for my present purposes these possibilities will be left alone.

believing all expressible propositions, *ex falso quodlibet*. In "Logic for Equivocators," David Lewis notes that by developing a fragmented model of agents' doxastic structures the formal epistemologist is able to bypass these issues, so long as his or her model meets four desiderata. Andy Egan's "Seeing and Believing" works towards a more fully developed theory of belief fragmentation while urging that fragmentation is a more accurate approach to conceptualizing cognition than non-fragmented models. An exceptionally sophisticated picture of belief structure fragmentation is provided by Samir Chopra and Rohit Parikh in "Relevance Sensitive Belief Structures," but Chopra and Parikh's model runs into plausibility concerns—or at least reveals room for improvement—with respect to its conception of belief revision. In order to avoid these concerns while retaining the benefits of Chopra and Parikh's model, I proposed the Single Fragment Model. Along with satisfying the basic desiderata set by David Lewis, the SFM is able to use the notion that certain beliefs are more or less central to agents' belief corpuses in order to provide a more psychologically plausible account of belief revision and querying than that provided by Chopra and Parikh's model. The upshot of focusing on the psychological plausibility of fragmentation models is that, given a sufficiently developed and psychologically realistic model, it is appropriate to assume that our doxastic systems are fragmented in some similar way, and we may acknowledge that whatever problems the model is capable of eliminating can be eliminated from our theories regarding the beliefs of actual agents. The SFM, as a psychologically plausible model of fragmented belief structures, allows us to escape the problem of being forced to commit any agent with incoherent beliefs to inadvertently believing all expressible propositions.

As a formal tool, the SFM adequately treats the notions that beliefs can "activate" and "deactivate," but the extent to which the illustration of cognition provided by the SFM accurately represents *memory* is not yet clear. In the following chapter I will shift my attention to what may be properly considered the philosophy of memory, with the aim of investigating the extent to which the SFM mirrors our actual theories of memory and the extent to which the deactivation and reactivation of beliefs affects their justificatory statuses.

Chapter 2

Memory Trace Theory and Justification Preservation

2.1 Reconstructive Memory

The notion of preservative memory, as characterized by Tyler Burge,[1] is memory that "preserves thoughts and their assertive mode, and does not contribute new elements in a justification, or add to justificational force," or, as he alternatively phrases his definition, "preserves beliefs with their justifications, but contributes no independent source of justification" (37). Consider, for example, an agent working on a proof of some philosophical principle p. This agent, on the basis of some set of established presuppositions and accepted rules for inference, works to prove that p is a consequence of her assumptions. At the end of the proof, the agent does in fact arrive at p as a logical consequence of her assumptions. The proof of p takes the agent all day, however, so the agent decides to wait until the following day to evaluate the implications of accepting p. According to Burge, if the agent's memory is preservative, then when the agent returns to work the day following her derivation of p her belief that p is in the same mode and justified to exactly the same degree and for exactly the same reasons that it was when she first finished her derivation. This is to say that the fact that the agent's present belief that p was produced by memory is wholly irrelevant for our understanding of the justificatory support for her present belief that p. When an agent employs a preservative memory faculty in recalling a belief, the recollected belief will be activated (in the language of the Single Fragment Model) with exactly the same justificatory support that it enjoyed the last time that it was active.

Given our frequent reliance on memory-beliefs through the courses of our thought processes it would certainly be ideal if our memorial faculties preserved justification through acts of recollection in this way, but it is possible to question whether or not human memory may be thought of as a preservative mechanism. This challenge is raised by David Christensen and Hilary Kornblith in their paper "Testimony, Memory, and the Limits of the A Priori,"[2] who argue that we do not employ preservative

[1] Burge, Tyler. 1997. "Interlocution, Perception and Memory." *Philosophical Studies* 86: 21-47.

[2] Christensen, David and Hilary Kornblith. 1997. "Testimony, Memory, and the Limits of the A

memory. According to Christensen and Kornblith, the notion of preservative memory requires an oversimplified "hall of records" conception of memory, in which "encoding is a matter of, in effect, taking a photograph of the passing scene; storage is a matter of placing that photograph in a file; and retrieval is a matter of taking the photograph out of the file"—a conception that is "fundamentally at odds with cognitive psychology"[3] (15). Contrary to the hall-of-records view, Christensen and Kornblith characterize the more psychologically sound picture of memory by reporting that "memory is far more constructive, and less passive, than such a picture would suggest. Our background concerns, interests and other beliefs – whatever their sources – affect the process at each of the three[4] stages," and thus we must take into consideration that "memories are often conditioned by, and in a sense incorporate, our background beliefs. Our remembering that P is supported by P's connections – its inferential connections – with our background beliefs" (15). The significance of this point for Christensen and Kornblith's argument is that if memory is understood as a reconstructive process in which the contents of memory-beliefs have to be rebuilt with the aid of connections to closely connected background beliefs, and the influence of one's background beliefs is liable to place further justificatory constraints on recollected beliefs, then the justificatory supports for our memory-beliefs will differ from their original counterparts in non-preservative ways. Christensen and Kornblith's essential intuition is that it is a necessary consequence of adopting a reconstructive picture of memory that agents' memory-beliefs will derive justificatory support from their background beliefs, and thus Burge's theory of preservative memory is inapplicable to our actual considerations of the transmission of justification through human memorial processes (16). In order to assess whether Burge's theory of preservative memory succeeds, we must first investigate whether Christensen and Kornblith's intuitions about reconstructive memory are correct.

Precisely how Christensen and Kornblith understand the reconstructive nature of memory and the "inferential integration" of memory-beliefs within background beliefs is not fully illuminated, but distinguishing between *semantic* memory and *episodic* memory may provide some light by which to examine their theory.[5] Semantic memories are, roughly speaking, the factive memory-beliefs that can be expressed by "remembers *that*" claims, such as "I remember that I need to bake a cake for Justin's birthday this year," whereas episodic memories are memories of "personally experienced events" such as, "I remember feeling ashamed on Justin's birthday last year when I forgot to bake him a cake" (Byrne, 16; citing neurobiologist Tulving[6]).

Priori." *Philosophical Studies* 86: 1-20.

[3]As a paradigmatic authority on the theory of memory provided by contemporary cognitive psychology, Christensen and Kornblith cite the following: Roberta Klatzky. 1975. *Human Memory.* San Francisco: W. H. Freeman and Company. They do not, however, cite any particular claims or passages.

[4]Christensen and Kornblith identify the three stages of memory as encoding, storage, and retrieval (15).

[5]These two types of memories are canonically recognized. See page 6 of: Matthen, Mohan. 2010. "Is Memory Preservation?" *Philosophical Studies* 148: 3-14; or page 16 of: Byrne, Alex. 2010. "Recollection, Perception, Imagination." *Philosophical Studies* 148: 15-26.

[6]Specifically, Byrne cites page 1506 of: Tulving, E. 2001. "Episodic Memory and Common Sense:

We may note that while semantic memories are capable of relating *a priori* knowable content, episodic memories are necessarily *a posteriori*—their content, by definition, only relates particular observations and experiences. Given that belief formation does not occur independently of our sense perceptions, emotions, or any other elements of our conscious experience during instances of belief formation, it is possible that the claim being put forward by Christensen and Kornblith is that all semantic memorial content must be embedded in the context of some or another memory of a previous unified conscious experience, and because such an experience must involve empirical information our semantic memories can only be accessed by recollecting an episodic memory and then somehow inferring the "purely semantic" content from the rest.[7] On this interpretation of Christensen and Kornblith, semantic memory-beliefs are only accessible by way of recalling certain episodic memories which contain those semantic contents, such that semantic memories are only accessible by way of an inference from the contents of particular episodic memories, and thus the memorial process places further justificatory constraints on one's memory-beliefs.

This seems an uncharitable interpretation of the sort of integration that Christensen and Kornblith are interested in, however, for the claim that all of an agent's memory-beliefs must be provided by way of making inferences from episodic memories seems unlikely. As a thought experiment, one might try recalling one's own name, paying attention to whether or not some particular experience of recognizing one's name (*e.g.*, looking at a name tag that says 'Alex') must first be recalled in order to recall *that* one's name is such-and-such. While such episodic memories certainly can be recalled, it does not seem plausible that one must recall such instances of recognizing one's name every time that one signs an email or fills out a form. It may be further remarked that while acts of engaging semantic memories may often produce mental experiences similar to certain past experiences (*e.g.*, encountering a mental image of a map of Europe when asked where Italy is located relative to Germany), it is doubtful that even these are necessarily episodic memory-beliefs that speak to any one past experience. It is possible that recalling a map of Europe as a mental image, for example, is simply an image-based semantic memory (thought of as "I remember that Europe looks like this") rather than an episodic memory of some actual map of Europe that one has previously seen ("I remember seeing that Europe looks like this"). That semantic memories are only accessible *via* inference from episodic memories is unsupported, and for this reason I will reject that Christensen and Kornblith's notion of the integration of particular memory-beliefs in background beliefs is meant to signify the integration of all semantic beliefs within episodic beliefs.

An alternative interpretation of Christensen and Kornblith's claim that all memories are "inferentially integrated" is that they mean to say that memories are packaged together, and that instead of individual sentence-like beliefs being recollected, memory involves the recollection of paragraph-like beliefs, such that in order to arrive at

How Far Apart?" *Philosophical Transactions: Biological Sciences* 356: 1505-1515.

[7]Because Christensen and Kornblith present their argument as a rejection of Burge's theory, I assume that they must be making a strong claim such as this—it is clear that Burge is not arguing that all memory is preservative, and thus his theory is not incompatible with the claim that some memory-beliefs are arrived at in this way.

any single semantic memory-belief requires an inference from the whole "memory-package" of which that individual belief is a part. Because such a "memory-package" is initially presented to a recollecting agent as a single memory-belief content, any component memory-belief content that is inferred from the memory-package would derive its justificatory status from the agent's justificatory relation to the entire memory-package, and thus the memorial process is seen to exert an influence over the justificatory statuses of agents' memory-beliefs. This interpretation of what Christensen and Kornblith intend to express by their claim that memory-beliefs are "inferentially integrated" with background beliefs enjoys support from its reflection in the following example, which Christensen and Kornblith use to illustrate how their notion of the "inferential integration" of memory-beliefs is problematic for Burge's notion of preservative memory (16-7).

The example begins by introducing two agents: Sam and Sophie. The first agent, Sophie, has a belief that the Vikings came to America before Columbus, but does not remember how she came to form this particular belief. This particular belief of Sophie's is connected to (*i.e.*, inferentially integrated with) a number of her other beliefs about the Vikings, all of which are well-formed and identifiable as having been acquired while she was attending an exceptional history course. The second agent, Sam, also believes that the Vikings came to America before Columbus. Like Sophie, Sam does not recall how he came to form this belief about the Vikings, and his belief that the Vikings came to America before Columbus is "inferentially integrated" with his other beliefs about the Vikings (16). Whereas Sophie's background beliefs were acquired while she was attending a history course, Sam's background beliefs are all based on the claims of a "crackpot history buff," which Sam received by reading the crackpot's book, "a work full of bad reasoning and unsupported conjecture, which claims that the Vikings invented the electric light bulb, that they discovered the vaccine against polio, and so on" (16). The key move in Christensen and Kornblith's example with Sam and Sophie is as follows:

> Now the reason that Sam believes that the Vikings preceded Columbus is not that the book says so. In fact, the book does not take any stand on the matter. But Sam remembers that the Vikings preceded Columbus only because this proposition is inferentially integrated with the ones he does believe on the book's authority. If it were not for its inferential connections with his irrational book-based beliefs, his memory that the Vikings preceded Columbus would have faded out long ago, and he would not have the corresponding belief today. In this case, it seems to us that Sam's belief that the Vikings preceded Columbus is not justified. (17)

Presumably, the same is taken to hold for Sophie, but the inferential connections sustaining her belief are justified. The intuition that Christensen and Kornblith put forward is that Sam's belief "The Vikings came to America before Columbus" is less justified than Sophie's content-identical belief—even if we assume that the two beliefs had the same justificatory statuses when they were acquired—because of the differences in the justificatory statuses of their Viking-relevant background beliefs.

According to the "memory-package" reading of Christensen and Kornblith's notion of inferential integration, the memory-belief content "The Vikings came to America before Columbus" is *only* accessible by Sam and Sophie in virtue of standing as a sentential proposition p_k in some Vikings-memory-belief-paragraph P_V.[8] Applying this to Christensen and Kornblith's example, then, we see that in order to recall their particular memory-beliefs that the Vikings came to America before Columbus, Sam and Sophie must infer that proposition from their respective belief-paragraphs. While the logical operations required to do this may be *a priori* warranted in certain cases, we may note that the inferred singular belief that the Vikings came to America before Columbus will not necessarily enjoy the same justificatory status that it did when that singular belief was first arrived at. Consider Sam's case, for example, in which the belief "The Vikings came to America before Columbus" may be assumed to have enjoyed a high degree of justification when Sam first believed it. According to the memory-package reading of Christensen and Kornblith's integration theory, Sam's memory belief that "The vikings came to America before Columbus" is only accessible by inference from a "memory-package" composed of Sam's other Viking-relevant beliefs, all of which were individually unjustified and thus endow the memory-package belief-paragraph with a low degree of justification. Because Sam's Viking-relevant belief-paragraph has a very low degree of justification, Sam's inference from the memory-package belief-paragraph to the individual proposition "The Vikings came to America before Columbus" will result in his individual belief having a low degree of justification, despite the fact that it can be traced back to a highly justified belief expressing exactly the same content. Even assuming that Sam can use an *a priori* warranted method of inference to arrive at his present belief that the Vikings came to America before Columbus, the requirement that Sam must infer this belief from an unjustified paragraph-like package of memory-beliefs necessarily makes his present belief unjustified (or at least significantly less justified than its initial counterpart).

What the Sam and Sophie example reveals is that if individual memory-beliefs must be inferred from memory-packages, then the justificatory statuses of the inferred beliefs will not be determined solely by the justifications which supported their original counterparts, but by the justificatory statuses of the complete memory-packages that they are stored within. We thus see that in the memory-package reading of Christensen and Kornblith, individual beliefs are not preserved with their justifications, and thus we are shown not to employ a preservative memory mechanism. This conclusion, that Burge's theory of preservative memory has been defeated, only follows from the "memory-package" interpretation of Christensen and Kornblith's commitment to the inferential integration of memory-beliefs if *all* memory-beliefs must be arrived at through these sorts of paragraph-sized memory-packages, however, and I do not see any reason to accept this condition. Just as one can recall one's name without recalling a particular episode of hearing or seeing one's name, it seems one can recall individual

[8]For the sake of tidiness one might think of P_V is a long conjunction of the form $P_V = p_0.p_1...p_k...p_n$, although we may note that most belief-paragraphs would likely be tied together with more complex connectors, such as "because," "therefore," etc.

beliefs without inferring their contents from a paragraph-sized block of other relevant beliefs. I can, for example, recall that $a^2 + b^2 = c^2$ without recalling any other basic facts about geometry. While I may not be able to recall *why* the Pythagorean Theorem is true without recalling facts about the relationships between the side lengths of rectangles and those rectangles' areas (and whatever other facts may be required to prove the Theorem), Christensen and Kornblith's claim hinges on my inability to recall that $a^2 + b^2 = c^2$ without first recalling a paragraph of geometrical facts which includes the Pythagorean Theorem as one element. This required inability makes the memory-package reading of Christensen and Kornblith's argument seem just as troubled as the previously considered theory that all semantic memory-beliefs must be inferred from episodic memory-beliefs—both defy intuition, and neither is given compelling evidential support through the course of Christensen and Kornblith's discussion. Ultimately, Christensen and Kornblith's rejection of preservative memory appears to lean on a vague appeal to the notion that memory is a reconstructive process (15), but in some way or another depends upon adopting an interpretation of this psychological commonplace that delivers implausible consequences. In order to assess whether or not memory can be preservative, then, we first need to understand what sense can be made of the claim that memory is a reconstructive process.

2.2 Memory Traces

Mohan Matthen considers an alternative conception of the reconstructive nature of memory in his paper "Is Memory Preservation?"[9] In "Is Memory Preservation?" Matthen introduces the notion of a *memory trace* as "the continuing state that is preserved in [one]" by memory, and distinguishes the memory trace from the belief that it corresponds to by emphasizing that a memory trace is "continuously occurrent and present" in an agent[10] (7). These traces are not themselves beliefs, since they persist through the activation, deactivation, and reactivation of an agent's memory-beliefs, and can be contrasted with the "dispositional, and only intermittently occurrent" beliefs that memory provides us access to (7). Importantly, Matthen holds that these memory traces can "[cover] semantic memory as well as episodic memory" (7). Matthen's memory traces allow for the possibility that a purely semantic memory-belief can be understood as the end of a reconstructive process that does not demand inference from background beliefs. That is, memory traces allow for beliefs such as "All crows are black" to be stored individually in memory, such that they can be reactivated in a reconstructive way without the need for inference from any supporting episodic beliefs. The sorts of inferences from background beliefs that appear to be demanded by Christensen and Kornblith's account are, under Matthen's memory trace theory, not seen as necessary steps in the reconstruction of both types of memories. Without further support, Christensen and Kornblith's argument that inference from background beliefs plays a necessary role in all recollection because memory is constructive comes up short, assuming we accept that it is something akin

[9]Matthen, Mohan. 2010. "Is Memory Preservation?" *Philosophical Studies* 148: 3-14.

[10]Matthen cites the use of similar notions in several other essays on the topic of memory (7).

to Matthen's memory traces which grounds the reconstructive aspect of memory.

The assumption that we employ something akin to memory traces may be called into question, but I believe that the notion of memory traces is tenable. Recall that the basis for theories of belief fragmentation is precisely that it has been found psychologically implausible to model actual agents as capable of simultaneously engaging all of their belief contents, and that we thus needed to introduce the distinction between agents' active beliefs and deactivated beliefs as a means of reducing the amount of belief content that agents can be held accountable for at any given time. If we understand belief as a cognitive relation to a propositional content, however, a "deactivated" belief is not properly speaking a belief at all, since even though an agent's mind may be disposed towards coming into contact with some propositional content that is similar (possibly identical) to previously considered content, we see that if whatever exists as a connection between prior and remembered belief contents is something other than mental state content, then by definition that connection cannot be a belief.[11] It is thus necessary to posit a certain kind of entity which is not itself mental state content but which is able to encode mental state contents, such that similar (ideally identical) mental state contents might be reproduced at a later time. Memory traces, then, are simply defined as these non-belief things which are disposed to produce mental state contents that are (a.) experienced as remembered and (b.) causally connected to previous mental state contents. Memory traces preserve previous mental state contents while being distinguishable *in kind* from the contents that they preserve, and thus allow for a reconstructive memory process that does not necessitate the kind of inferential integration Christensen and Kornblith believe reconstructive theories of memory require. This minimally specified notion of memory traces appears necessary for any theory of the memorial process other than a literal "Hall of Belief Contents" picture, and thus I will assume that we employ something akin to memory traces. Finally, given that the notion of memory traces is compatible with individual beliefs being stored and reactivated *via* memory traces, it is clear that Christensen and Kornblith's inferential integration argument is not entailed simply by the fact that memory is a reconstructive process.

While something akin to Matthen's memory traces is sufficient for beliefs to be encoded, stored, and retrieved by a reconstructive memorial process—Christensen and Kornblith's three desiderata (15)—Matthen ultimately finds them problematic for the theory that the causal histories of memory-beliefs preserve their justificatory status, arguing, "Belief or experience itself is not preserved in memory, nor is 'representational content' . . . Nor does the warrant of my original belief survive intact" (5). In short, Matthen's argument for this claim is that when memory traces produce mental state contents, the content of the mental state is "experienced as having occurred before" (9), such that memory traces may be thought of as adding a "tense operator" onto the beliefs they preserve as "a signature of the memory system," thus making the content of memory-beliefs necessarily distinct from the contents of their original counterparts (11). According to Matthen, we cannot recall a memory-belief

[11] While this claim appears to raise a serious problem for the SFM, I will explain later in this chapter that any apparent conflict between this claim and the SFM are merely superficial.

without engaging some amount of extra content added by the tense operator. The necessity of this additional memory trace signature content is particularly palpable when considering cases of episodic memory. When I recall the experience of walking along the Thames River, for example, I do not become confused about whether I am strolling through London or sitting in my living room. I recognize that my memory of being in London is a past experience in virtue of the means through which that content is presented. More generally, the story Matthen's theory tells is that if an agent affirms some content c—and thus, in the language of the SFM, introduces c into her belief corpus—then what becomes stored in memory is a memory trace that would reactivate with a content along the lines of "Previously(c)," from which one might go on to infer c (depending on the particular c).

Simply speaking, memory traces may be understood as mental state content inducers; they persist within the mind and are disposed to provide agents with access to mental state contents that are similar to contents of previous mental states (identical, when memory functions ideally) but which are not themselves mental state contents. Furthermore, given that the contents induced by memory traces are assumed to carry some form of additional signature content in virtue of the memorial process that produced them, their justificatory statuses seem liable to differ from the justificatory statuses of their original counterparts, and thus memory trace theory threatens the possibility that we employ a preservative memory mechanism.

Concluding his paper, Matthen writes, "It is commonly held that memory is preservation, and surely it is. But it is wrong to think that memory is a preservation of what is experienced or represented in the memory-experience—an image or a belief. What is preserved is a trace from which it is possible to reconstruct an image or belief" (14). Component contents of our mental states are only preserved by memory in virtue of the capability of corresponding memory traces to induce similar mental state contents at a later time. Reactivated mental state contents necessarily fail to be identical to their original counterparts in virtue of the additional tense operator content that is produced by the reactivation of a memory trace. "The endpoints of the memory process may be similar to one another," Matthen points out, "But this should not lead us to think that there is a single process of preservation in which the interior points are the same as the endpoints" (13). While accepting the notion of memory traces allows for a reconstruction-based picture of memory that does not depend on the sort of inferences that Christensen and Kornblith believed any reconstructive conception of memory would require, the notion of memory traces does create further issues for a positive theory of how memory preserves justification. Preservative memory requires that memories "preserve thoughts and their assertive mode" (Burge, 37), but memory traces are only capable of approximating these features. Matthen admits that he is uncertain about how significant of a correction the notion of memory traces demands from Burge's view, but Matthen is certain that some adjustment will be required if we wish to say that actual agents enjoy a preservative memory mechanism (14). In the following sections, I will examine how the additional content that memory traces add to reactivated beliefs impacts the justificatory status of those beliefs, with the aim of providing a satisfying response to Matthen's uncertainty about the degree to which memory trace theory conflicts with

Burge's notion of preservative memory.

2.3 Memory Traces and the Memory Justification Principle

The primary problem that memory trace theory makes for the notion of justification-preserving memory is that memory traces interrupt the causal histories of memory-beliefs, which is problematic because it is the causal histories of beliefs which are commonly thought to track the flow of justification through inferential sequences.[12] Memory traces obscure the causal histories of beliefs, since what is preserved in memory according to memory trace theory is not actually mental state content, but some other thing that is able to approximately reproduce previously engaged mental state content (14), while also contributing additional content in the form of the memory trace signature. Because it is intentional attitudes (belief, desire, etc.) towards propositions that are said to be justified or unjustified, justification cannot be attributed to memory traces themselves. How, then, might we understand the justificatory statuses of the mental state contents produced by the activation of memory traces?

First, we might recall that the theoretical utility of memory is that it helps cognitively limited agents such as human beings behave successfully by providing us with information that is similar to information which we have previously engaged with, providing us with this information when it is relevant to our present cognitive situation, and freeing up cognitive resources by storing that information in a non-contentful form when it is not relevant. The SFM characterizes this as a relevance querying process based on similarities between the sub-propositional constituents of inactive beliefs and some query. The reason why the SFM uses such a fine-grained basis for querying (at least compared to the proposition-level querying of the *B*-structures model) is that the ways by which actual agents' memory mechanisms are able to draw associations between present mental state content and prior content can be remarkably nuanced. Regardless of what it is that persists through belief deactivation,[13] it is clear that our actual memory processes are capable of providing us with content that is subtly relevant to our present cognitive states. Given that the type of relevance querying that takes place in actual agents can use the subtle similarities between present and previous contents as the basis for content reactivation, the properties of memory traces must be such that they precisely reflect the details of prior mental state contents in order for this sort of querying to take place. What this shows is that while memory traces are not themselves beliefs, their structural features unmistakably correspond to the contents of the mental states that they are

[12] Momentarily disregarding considerations of memory, I believe that most philosophers are roughly inclined to think that the justificatory statuses of beliefs may be examined by considering the facts surrounding how those beliefs were acquired and maintained, which is to say that this assumption that the causal histories of beliefs track their justificatory statuses should be relatively uncontroversial.

[13] For the sake of brevity, from here on out I will borrow Matthen's terminology and call these objects *memory traces*.

determined by and are disposed to approximately reproduce upon activation.

For the sake of having a way to refer to the process by which a memory trace comes to reflect a particular content, let us say that a memory trace was *sculpted* with respect to some mental state iff that mental state uniquely caused the memory trace to become able to produce mental state content by activating. Generally speaking, we see that mental states exert a causal influence on the structure of memory traces, and that these structural properties of memory traces cause the production of particular mental state contents during reactivation. That is, memory traces are sculpted with respect to mental states, and through this sculpting process memory traces become capable of producing mental state contents that are similar[14] to those which were represented by the mental states with respect to which those memory traces were sculpted. Furthermore, the content of the belief that a memory trace activates will, under normal circumstances, approximate the content of the mental state that initially sculpted the memory trace (momentarily setting aside the matter of the memory trace signature content). While content is not necessarily perfectly preserved[15] by memory traces, the mental state contents produced by the activation of memory traces are undeniably[16] causal successors of the mental states with respect to which those memory traces were sculpted.

In order to determine the justificatory statuses of memory-beliefs, we will need to investigate the extent to which the contents produced by the activation of memory traces inherit the justification from the state with respect to which those traces were sculpted. Given that the "flow" of justification is typically traced through the causal histories of beliefs, and that the content provided by reactivated memory traces is causally influenced by the content of the mental states that originally sculpted those memory traces, our account will have to be at least partially based on considerations of the extent to which the content produced by an activated memory trace is similar to the content of the mental state with respect to which that memory trace was sculpted. Our account will also have to take into consideration the fact that the activation of memory traces is thought to produce contents that include distinctive memory trace signatures. With this in mind, let us consider the following principle:

> *Memory Justification Principle:*
> Let s_1 be a mental state with respect to which a memory trace m is sculpted at time t_1. Let c_1 be the propositional content represented by s_1, and let j_1 be the justificatory status of c_1 at t_1. Let s_2 be the mental state produced by the activation of m at some time $t_2 > t_1$. Let c_2 be the

[14]Recall that because of the addition of the tense operator signature the content produced by an activated memory trace cannot be identical to the content of the state with respect to which it was sculpted.

[15]Imagistic memories frequently serve as paradigmatic cases of imperfect memory preservation. Very few individuals, for example, could number how many stones compose Stonehenge from imagistic memories, while many people can easily recollect what Stonehenge looks like.

[16]While there are skeptical scenarios such as the five-minute world (where the world only came into being five minutes ago, and all of our memories of prior events came pre-packaged) and instances where people might feel that they remember events that never happened, I do not take these to be cases genuinely involving reactivated memory traces.

> propositional content represented by s_2. Finally, let the operator $sig()$ be the memory trace signature included in c_2. We say that the justificatory status j_2 of c_2 at t_2 is given by $j_2 = p * j_1$, where p is the proportion of c_2 that is also content of $sig(c_1)$.[17]

The essential idea behind the Memory Justification Principle is that the degree to which a previous mental state has been preserved through the memorial process corresponds to the degree to which the reconstructed mental state's propositional content is identical to the propositional content represented by the original state, with the caveat of having to account for the memory trace signature adding some amount of content to one's recollections. The degree to which remembered content is identical to the content represented by the mental state with respect to which the relevant memory trace was sculpted is thought of as reflecting the degree to which the remembered mental state causally succeeds the original mental state, and, by extension, reflects the degree to which the content of the recollected mental state causally succeeds the content of the original mental state. Because justification tracks the causal histories of mental state contents, the degree of sameness of content between the relevant mental states should correspond to the degree of sameness of their justificatory statuses, in virtue of representing the degree to which the features of the original mental state causally influenced the features of the recollected mental state. Furthermore, because it is counterintuitive to think that a recollected mental state content can be more justified than its original counterpart,[18] it is worth noting that one important feature of the Memory Justification Principle is that it sets an upper bound on the justificatory status that a remembered mental state content can enjoy which is equal to the justificatory status of the content of the mental state with respect to which the relevant memory trace was sculpted—a bound which can only be met when the memory trace signature provides the only content represented by the remembered mental state which was not also represented by the original mental state. For the sake of establishing more concrete support for the Memory Justification Principle, I will focus the remainder of this section on illustrating how the Principle functions through a series of examples, with the intention of showing that it provides an intuitive measure of the degree to which memory preserves the justificatory statuses of mental state contents.

First, consider an agent who has a highly justified belief that most squirrels are brown and sculpts a memory trace with respect to a mental state representing that most squirrels are brown. At some later time that memory trace is activated and produces a content along the lines of "previously(most squirrels are brown)," assuming for our present purposes that the memory trace signature is the past-tense tense operator "previously." We can add the same tense operator to the original belief content,

[17] As an extremely oversimplified example, if one thinks of propositional content in terms of possible worlds and assumes that there are a finite number of possible worlds, then p may be calculated in the following way: Let W_1 be the set of possible worlds selected by $sig(c_1)$. Let W_2 be the set of possible worlds selected by c_2. Let n equal the size of the intersection $W_1 \cap W_2$. Finally, let m equal the size of W_2. In this case we would say that $p = n/m$.

[18] In the following chapter I will discuss the possibility of remembered mental state contents being more justified than their original counterparts in greater detail.

and see that the two contents are identical. In such a case, memory has ideally approximated content preservation, and thus the Memory Justification Principle allows us to treat justification as having ideally "flowed" through the causal mechanisms of the agent's memory. This is the result that we would hope for, as agents who frequently rely on memory.

Memory does not always function ideally, however, so we may consider the situation where the memory trace from the previous example activates, but produces the content "previously(some squirrels are brown)." While memory has not perfectly replicated the original belief content, the content that the memory trace produces is a weaker sub-content of "previously(most squirrels are brown)," and for this reason the content produced by the memory trace still enjoys the same justificatory status as the original content. This is an intuitive result for two reasons. First, we can note that the weaker content, though not identical to the tense-adjusted original content, is implied by the original content and thus content has been preserved through the agent's memory mechanisms, albeit only partially. Second, we might suppose that if the agent had originally believed "some squirrels are brown" instead of "most squirrels are brown," his belief would have been more justified, given that he has not seen most squirrels but has seen some squirrels. Because the weaker content is produced by memory, however, the agent's memory-belief "previously(some squirrels are brown)" cannot be more justified (*qua* memory-belief) than the content that would have been preserved by an ideally functioning memory. The Memory Justification Principle thus provides an intuitive result in situations of partially preserved contents.

Next, we may consider a situation in which this same memory trace produces the stronger content "previously(all squirrels are brown)." In this situation, we must consider the degree to which the content of this stronger proposition is not the content of "previously(most squirrels are brown)." While there are a variety of approaches to approximating the difference in content between these two propositions, I will for the ease of illustration use the following oversimplified method: assuming that there are a finite number of worlds, take the number of possible worlds in which all squirrels are brown divided by the number of worlds in the union of the set of possible worlds in which all squirrels are brown and the set of possible worlds in which most squirrels are brown. Let us suppose for simplicity that our calculation yields the result that only 20%[19] of possible worlds that are all-or-mostly-brown-squirrels worlds are all-brown-squirrels worlds. In this case, the content "previously(all squirrels are brown)" enjoys a justificatory status equal to 20% of the original belief's degree of justification. Justification still "flows" causally through the memory mechanism, but because memory has added extra information to the original content, the belief produced by the memory trace is significantly less justified. Again, the results of applying the Memory Justification Principle are in agreement with what I believe to be reasonable intuitions.

The previous three cases have focused on the Memory Justification Principle's

[19]The use of numerics in these examples is simply to make my illustrations more transparent. Substituting approximate terms such as "many," "most," "a few," etc. only impedes the application of the Memory Justification Principle insofar as it demands reflection upon how much justification is affected when remembered content is "mostly similar" instead of "80% similar"

application to semantic memories, but the Principle also applies to episodic memories. Suppose that I see a black squirrel while sitting on the front lawn of Reed College. My mental state at that time contains the perceptual content of not only the squirrel, but also contains perceptual contents corresponding to the trees around the perimeter of the lawn, the east-facing wall of MacNaughton, the other people enjoying the lawn, etc. I assume that in this situation the content of my mental state is highly justified—I have not ingested any powerful hallucinogens, I am feeling wide awake, etc. Further, suppose that a memory trace is sculpted with respect to this perceptual state. I may, at a later time, activate this memory trace and produce a content that is equivalent to the content of seeing a black squirrel, plus the tense operator. I may report this content by saying "Previously I saw a black squirrel." While I may be unable to recount how many other people were on the lawn, or exactly how far away I was from the trees, it seems, intuitively, that I ought to be highly justified in my report "Previously I saw a black squirrel." The Memory Justification Principle delivers this result; the situation is analogous to that in which I reconstructed a belief content weaker than the one that I originally maintained. Because my perception of the black squirrel (as one object among other things) corresponded to a highly justified mental state, my recollection of seeing a black squirrel is also highly justified, even though I cannot recall the other things that were around the squirrel—the content of the reconstructed state is more or less exactly content that was part of the original state, and thus justification is essentially preserved according to the Memory Justification Principle.

As the preceding cases show, the Memory Justification Principle delivers intuitive results regarding the preservation of justificatory status through a picture of memory based on something akin to Matthen's memory traces. This only accounts for the justificatory statuses of recollected mental states whose contents still include the tense operator supplied by the memory trace signature, however, and it has not yet been shown that justification can be preserved through the inferences necessary to eliminate tense operators.

2.4 Tense Operator Elimination

Suppose that I have previously proven the Pythagorean Theorem, and that I have a memory trace that was sculpted with respect to my belief that for any right triangle $a^2 + b^2 = c^2$. When this memory trace activates it will, supposing that content has been ideally preserved, roughly produce the mental state content "$prev(a^2+b^2 = c^2)$," where *prev* is the general past-tense tense operator. If I need to use this memory to calculate the hypotenuse of a triangle right now, however, it does me no good to know that the Pythagorean Theorem *was* true; I need to eliminate the tense operator from the content of my memory-belief, and this will require an inference of some kind. In order to achieve the sort of truly preservative memory that Burge endorses, it is necessary that the tense operator elimination inference can preserve *a priori* warrant for content. The question I will address in this section is what conditions allow for *a*

priori warranted elimination of the past-tense operator *prev*.[20]

We may begin our considerations by introducing the notion of *eternally true* propositions as propositions that are true at all times iff they are true at any time. Paradigmatic examples of eternally true propositions are mathematical and logical truths. By definition, it is true that if a proposition p is eternally true, then the application of any tense operator to p is truth-preserving, since the application of a tense operator simply asserts that p is true at some proper subset of the set of all times, and it follows analytically from the fact that p is true at all times that p is true at any proper subset of the set of all times. It also follows from the definition of eternally true propositions that if an eternally true proposition has had a tense operator applied to it, then the elimination of that tense operator is truth-preserving, since the tense operator asserts that the proposition it is operating on is true at some time, and an eternally true proposition is true at all times if it is true at some time, and in virtue of the proposition's being true at all times it follows that it is presently true. More formally, the foregoing informs us that the following is *a priori*: $eternallyTrue(p) \Rightarrow (tensed(p) \iff p)$, and because *tensed* can stand for any tense operator, this also gives us $eternallyTrue(p) \Rightarrow (prev(p) \iff p)$, which is to say that it is *a priori* that if a proposition is eternally true then its having been previously true implies its being presently true. If the content produced by a memory trace is equivalent to the tensed proposition *prev(p)* and if p is an eternally true proposition, then because we have an *a priori* warranted inference from $prev(p)$ to p we have an *a priori* warranted method for eliminating the memory trace's tense operator signature. Thus any remembered eternally true proposition can have the tense operator contributed by the memory trace eliminated by a justification-preserving *a priori* inference.

As a more general case, we may define a *forever-after* proposition with respect to time t as a proposition that is true necessarily[21] at all times after t iff it is true at t.[22] Forever-after propositions are those propositions which cannot be made false after being made true. The class of forever-after propositions thus includes the set of eternally true propositions as a proper subset,[23] but is more general than the class of eternally true propositions in virtue of also including propositions which have not been true at all times. For example, "George Washington served as the first President of the United States" is a forever-after proposition with respect to some time on April 30th, 1789, since for all times after the first moment where George Washington served as the first President of the United States it cannot be made false that he served as

[20]More rigorously, we can define the past-tense tense operator *prev* by saying that for some time t_1 and set of times $T = \{t | t < t_1\}$, $prev(p)$ is true at t_1 iff p was true at some time $t \in T$.

[21]It could be, for instance, that between now and the time that the universe ends there is a red Toyota Avalon parked in my driveway, but this proposition would not count as a forever-after proposition, since it is only a contingent fact that some such car remain parked in my driveway for all times after the present moment.

[22]I use 'times' as a shorthand for "discrete time-slices of worlds."

[23]We can see that the set of eternally true propositions are a subset of ever-after propositions by noting that eternally true propositions are simply those propositions which are forever-after propositions with respect to the time $t_e \in T$, where T is the set of all times at the relevant world and t_e is the earliest time in that world.

the first P.o.t.U.S. Intuitively, then, whenever a true forever-after proposition has a tense-operator attached to it, the tense-operated proposition implies the present-tense forever-after proposition. This is to say that p follows *a priori* from $prev(p)$, so long as p is a forever-after proposition with respect to some past time.[24] We thus see that tense operators can be eliminated *a priori* at least in the case that the tense operator behaves like *prev* and is operating on a forever-after proposition with respect to some past time. Therefore, so long as what memory exposes us to is propositions of the form $prev(p)$, then it is possible for *a priori* justification to be transmitted through the memorial process.

2.5 Generalized Memory Trace Signature Elimination

While the tense operator approach to understanding the memory trace signature would certainly be convenient as a means for supporting Burge's notion of preservative memory, Alex Byrne[25] finds the theory that the memory trace signature is simply a tense operator to be "mysterious" and "a dubious invention" (23). What Byrne finds troublesome about conceptualizing the memory trace signature (or, to use the term Byrne borrows from Holland,[26] "the Memory-Indicator"; 24) as a tense operator is that he believes that memories carry a "feeling of familiarity," which is "uncontroversially a feature of ordinary experience, [which] *already* carries the sense of pastness with it" (23). It is unquestionable that memories present themselves as being familiar, according to Byrne, but because familiarity implies past experience we are confusing an inference from the sense of familiarity that accompanies memory-contents to their being previously true with what is actually presented during acts of recollection—an experience whose content includes an explicit representation of something as being past. Byrne is not satisfied with the theory that the memory trace signature is some generic feeling of familiarity either, however, for a sense of familiarity does not by itself signify that a memorial process has been engaged, and one is (at least in most cases) able to phenomenally distinguish a mental state content produced by memory and contents produced by other faculties, such as imagination and perception. If I frequently imagine what it would be like to fly, for example, then I might experience a feeling of familiarity when imagining looking down through the clouds on Portland, but despite this feeling of familiarity I am able to recognize that this is an imagining and not a memory, even without performing any reasoning about my lack of super-powers. So long as a feeling of familiarity is not sufficient to recognize that the content of one's present mental state was produced by memory, then the content of the memory trace signature must be something other than a feeling of

[24] It may be possible that there are propositions for which the $prev(p) \Rightarrow p$ inference is *a priori*, but for my purposes it is sufficient that at least in the cases of forever-after propositions this inference is *a priori*.

[25] Byrne, Alex. 2010. "Recollection, Perception, Imagination." *Philosophical Studies* 148: 15-26.

[26] Byrne specifically cites pages 465-6 of: Holland, R. F. 1954. "The Empiricist Theory of Memory." *Mind* 63: 464-486.

familiarity.

Instead of viewing the memory trace signature as explicitly expressing pastness or familiarity, Byrne characterizes the memory trace signature as distinctive from other experiential modes only insofar it indicates that the content it operates on has been produced by memory—all we know is that it is "a sign or mark that enables one to know that one is recalling, and so not imagining" (24). We may then think of Byrne's memory trace signature as something along the lines of a *recall* operator, which only necessarily expresses the content that whatever it is operating upon has been produced by memory.[27] While the content contributed by this signature includes the "uncontroversial" feeling of familiarity that comes with memory, Byrne seems to take this as a sense of familiarity that is distinctly identifiable as the familiarity of a recollected content.

An intuitive advantage that the *recall* operator enjoys over the *prev* operator is that the *recall* operator helps explain why we frequently perform the cognitive shortcut of inferring a proposition's present truth from our recollection of its prior truth—because the content that the *recall* operator works on may remain entirely presently tensed, we see that the cognitive shortcut from the content "Recall that $a^2 + b^2 = c^2$" to "$a^2 + b^2 = c^2$" is intuitively easier to make than the shortcut from "Previously $a^2 + b^2 = c^2$" to "$a^2 + b^2 = c^2$," since no considerations regarding the differences in the states of affairs of different time-slices of the world are explicitly demanded.[28] Given that we infrequently expend any conscious cognitive energy on inferring that the content of what we recollect is still the case (especially in cases of eternal and forever-after propositions), one can see that there is a straightforward argument that the *recall* operator approach to the memory trace signature fits with our experiences of recollection slightly better than the *prev* operator. We will thus have to investigate whether it is possible for justification to be preserved through our inferences from $recall(p)$ to p.

We may note that an essential part of the *recall* operator approach to understanding the memory trace signature is that the content of the *recall* operator includes a memory-specific sense of familiarity, and, as Byrne points out, this feeling of familiarity "*already* carries the sense of pastness with it" (23). Even more explicitly, Byrne notes, "That an event was in the past is just another piece of information, and so can be part of the content of the recollection" (24). It is unclear, however, where this particular piece of information fits into Byrne's understanding of the phenomenology of memory-beliefs. Insofar as Byrne has characterized the *recall* operator method

[27] It is important to note that the fact that the *recall* operator only necessarily expresses that its content has been produced by memory, it is not necessary that it only expresses that the content it operates on has been produced by memory.

[28] Because the feeling of familiarity that comes with the *recall* operator is still taken as enabling us to make claims about the past (insofar as having a sense of familiarity requires recognition of the fact that one's present experience is similar to some past experience), and thus weakly implies a sense of pastness, we see that a thoroughgoing elimination of the memory trace signature will still require tense operator elimination. The significance of the ease of the cognitive shortcut from $recall(p)$ to p is just that in many of our actual experiences we do not immediately recognize that the propositions we remember (especially eternal propositions) are being expressed to us as having been *previously* true, and thus *recall* fits our experience of recollecting better than *prev*.

of approaching the content of the memory trace signature, we see that $recall(p)$ is essentially analyzable in the following way: $recall(p) = \exists e(e\ MemoryPresents(p))$, where e is a phenomenal experience and $MemoryPresents$ is the distinct mode of experiential presentation associated with memory which involves the sense of familiarity that is specifically characteristic of memory. The problem this presents is that given Byrne's rejection of the tense operator approach to understanding the memory trace signature, he would seem unable to accept that $e\ MemoryPresents(p)$ analyzes to $e\ Presents(prev(p))$, since this is just to say that $MemoryPresents$ is distinguished from "generic" phenomenal presentation by providing a sense of pastness, which Byrne is unwilling to accept. In keeping with Byrne, then, we must preserve $MemoryPresents$ as an unanalyzable feature of the experience e. This means that in cases where a sense of pastness is included in the content of a memorial experience expressible as $\exists e(e\ MemoryPresents(p))$, this sense of pastness must arise partly in virtue of the sub-content p, rather than solely in virtue of the $MemoryPresents$ experiential mode.

A further concern is that if we agree that whenever we are presented with a sense of pastness in our recollections, the analysis of our recollection's content must follow along the lines of $recall(p) = \exists e(e\ MemoryPresents(prev(p')))$, where p' is the untensed propositional content of p, then the Memory Justification Principle delivers the troubling result that all of our recollections that come with a sense of pastness are significantly less justified than the contents of the mental states with respect to which they were sculpted. Given that the memory trace signature is just the $MemoryPresents$ operator on this account, what the Memory Justification Principle would have us compare (in order to calculate the memory's degree of justification) is the proportion of content shared between the original mental state content and $prev(p')$. Even if we assume that the original mental state content was p', it would seem that the $prev$ operator sufficiently saturates the content of what it operates on such that p' and $prev(p')$ will turn out expressing largely different propositional contents, and thus the proportion of propositional content shared between the original mental state and the recollected mental state will be significantly less than 1, forcing a noticeable decrease in justification. So long as we wish to maintain (1.) that memories reasonably preserve justification, even in cases where they establish a sense of pastness, (2.) that the Memory Justification Principle is a roughly correct characterization of the extent to which justification is transmitted through memory, and (3.) that the memory trace signature is not simply a tense operator, then we must develop a plausible alternative analysis of $\exists e(e\ MemoryPresents(p))$ that is both compatible with the possibility of justified recollections and capable of explaining the sense of pastness contributed by the $MemoryPresents$ operator in this approach to conceptualizing the memory trace signature.

To begin developing such an analysis of the $\exists e(e\ MemoryPresents(p))$ characterization of recollection I will first focus only on episodic memories, such as when I recall seeing a squirrel at a time when I am not looking at a squirrel. Analyzing my experience of seeing a squirrel as having the content $\exists e'(e'\ PerceptionPresents(s))$, where s is a propositional content along the lines of "There is a squirrel" and e' is a quantification over the visual experience whose content is equivalent to s, we see that the

experience of remembering seeing a squirrel can be approached as the nested experiential existence claim $\exists e(e\ MemoryPresents(\exists e'(e'\ PerceptionPresents(s))))$. Given that we are capable of distinguishing the content of our present perceptions from recollected perceptions in virtue of the meta-content provided by the $MemoryPresents$ operator, then so long as we take the content presented within the $MemoryPresents$ context to be veridical we must understand that the perceptions we are exposed to within this context are not present,[29] and thus presumably past.[30] In this way, when I am recalling the squirrel I saw, my awareness that the content of that perceptual state is now being presented within a $MemoryPresents$ context saturates the perceptual content, such that in order to maintain that the information presented to me by both memory and perception truthfully represent the world, I have to acknowledge that what is presented within the $MemoryPresents$ context does not report what is presently the case, but rather as preserved information about what was once the case. I may thus infer $prev(\exists e'(e'\ PerceptionPresents(s)))$ from the content of the recollected image of the squirrel, since this appears the best way to reconcile the veridical presentation of my memory with the fact that it contradicts my present perceptual situation. More generally, we may observe that in any case of episodic memory, our experience of recollection will similarly warrant the formation of the belief that what our memorial experience presents is a proposition about the past, such that we may infer $prev(c)$ from $\exists e(e\ MemoryPresents(c))$ whenever c is a further existence claim about an experience. Furthermore, because the $MemoryPresents$ operator is taken to be universally operative through our memory experiences, we may make the same inference in cases where the content of a memory experience is semantic,[31] since this mode of presentation (*i.e.*, presentation of contents within the $MemoryPresents$ context) may be thought to generally provide us with previously engaged mental state contents. We are thus able to see that for any proposition p the following holds:

$$recall(p) = \exists e(e\ MemoryPresents(p)) \Rightarrow prev(p)$$

This analysis of the *recall* operator is precisely what we desired. We understand the generalized memory trace signature (*i.e.*, the *recall* operator, which is equivalent to a quantification of an experience of applying the $MemoryPresents$ operator to a content) as warranting belief formation about what was previously the case, but not as itself presenting pastness as content, thus satisfying Byrne's desideratum that

[29] As we may recall, Byrne's understanding of the phenomenal content of the memory trace signature involves a distinct feeling of familiarity, which implicitly "carries the sense of pastness with it" (23). My argument, to frame it as a response to Byrne's reliance on the "feeling of familiarity," is that the $MemoryPresents$ operator captures something like a feeling of familiarity by simultaneously presenting what it operates on to be true, while allowing for that content to disagree with the content of one's present perceptual state.

[30] In order to argue that "not present" must be read as "past or future" rather than simply as "past" in this context, one would have to argue that the experience of soothsaying is phenomenally identical to recollection. I have not yet seen a successful argument to this effect, and thus assume that in cases of $MemoryPresents$ experiences, "not present" is equivalent to "past."

[31] There may be other ways of eliminating the meta-content that the content is being presented by memory, such that we can immediately infer p from $\exists e(e\ MemoryPresents(p))$ for certain semantic memories. I privilege the inference method that I discuss simply because of its broad applicability.

the memory trace signature not be merely reducible to the application of a tense operator. Furthermore, because there is no tense operator saturating the content inside the memory trace signature, the Memory Justification Principle still allows for the possibility that the content of our recollections can be identical to the content with respect to which the relevant memory traces were sculpted (provided an appropriate application of the memory trace signature to the original contents), and thus allows for the preservation of justification from a mental state representing p to a mental state representing $\exists e(e\ MemoryPresents(p))$. In order for Burge's thesis that we employ a preservative memory mechanism to succeed, however, it would appear that we need to be capable (at least in some situations) of preserving justification all the way through—that is, there need to be situations in which justification can be preserved from $\exists e(e\ MemoryPresents(p)) \Rightarrow prev(p) \Rightarrow p$. This is the topic that I will turn my attention to in the following section.

2.6 Is Memory Trace Signature Elimination *a priori* Warranted?

We have already seen that there are certain situations in which p follows *a priori* from $prev(p)$, and now that we have seen that there is a general means for inferring $prev(p)$ from $recall(p)$ we can see that our verdict on whether or not Burge's notion of preservative memory succeeds will hinge on whether or not the aforementioned inference from $recall(p)$ to $prev(p)$ is *a priori* warranted. Because we have to infer $prev(p)$ from $\exists e(e\ MemoryPresents(p))$ on the basis of our experiential familiarity with the $MemoryPresents$ operator, it appears possible that this additional step may necessarily impede the flow of justification and accordingly make it impossible for memory-beliefs to enjoy *a priori* justification. We must consider whether the necessarily *a posteriori* basis for our inferences from $\exists e(e\ MemoryPresents(c))$ to $prev(c)$ forces us to reject the possibility that we can enjoy *a priori* warranted memory-beliefs.

Recall that the basis for our inference from $\exists e(e\ MemoryPresents(p))$ to $prev(p)$ was that we can observe that the experience involved in $MemoryPresents$-presentation is distinguishable from other sorts of experiences in a way that requires that what $MemoryPresents$ presents must be past, and thus depends on our *a posteriori* awareness of the phenomenal differences between certain kinds of experiences. The concern for the possibility of preservative memory, then, is that this *a posterioricity* demands that the inference from the experience of recollection to a belief about what was previously the case is empirical, and thus cannot transmit *a priori* justification. To counter this concern, we may first note that even our most rational thoughts involve some sort of *a posteriori* phenomenal experience. When a homework set for logic class tells us to apply *modus ponens* to the assumptions A and $A \Rightarrow B$, for example, we recognize that B follows by "seeing" how the rule applies in this situation, and thus our belief "B follows from $A \wedge (A \Rightarrow B)$" may only be achieved by having a particular experience. We do not, however, say that all of our beliefs are arrived

at by an empirical process because we can only believe propositions by standing in experience-based belief-relations to them. Rather, we distinguish empirical processes as those which allow us to examine the properties of objects using sense perception. While memory traces may possibly be thought of as existing outside of the mind, it is wrong to say that they have *as properties* propositional contents like "That I was sitting on the lawn and saw a squirrel" or "That the length of a right triangle's hypotenuse is equal to the square root of the sum of the squares of the side lengths." When a memory trace represents to me that I saw a squirrel, then, I am not using some perceptual apparatus to inform me of the properties of the memory trace—at best these properties may be inferred by working backwards from the contents that they produce, and even then only by reference to those contents (*e.g.*, being such that it can produce a recollection of seeing a squirrel). Insofar as we are performing an observation during memory trace activation, then, we are observing the reappearance of previously entertained mental state contents, rather using our sense perceptual faculties to examine something that exists outside of the mind. The fact that the contents of our recollections are presented to us by some memorial mechanism does not entail that we are provided those contents by an empirical process, and thus our exposure to the experiences quantified over in this analysis of the *recall* operator should not be understood as a kind of empirical information.[32]

Similarly, I do not see any reason to believe that the knowledge required to make the temporal inference from $\exists e(e\ MemoryPresents(p))$ to $prev(p)$ is an exercise of empirical knowledge—that is, an exercise of knowledge acquired through the use of our sense perceptual faculties. While the recognition of this distinction between memory-presented contents and contents presented through other experiential modes is necessarily *a posteriori*, it is a distinction based on the fact that memory informs us of the existence of experiences that are not in themselves (*i.e.*, independently of the $MemoryPresents$ experiential context) part of what we recognize as the contents of our present experiences. This is to say that at any given time when an agent has an experience of having content presented by memory, this experience is recognizable as a different type of experience from the other sorts of perceptual and doxastic experiences establishing that agent's unified conscious experience at that time. When memory veridically reports the existence of experiences that do not agree with the contents of an agent's other present experiences, then provided that we understand time as necessarily experienced in a progression from earlier to later, we must recognize that the contents of our memory experiences are recreations of previously entertained mental state contents. Despite the fact that recognizing the distinction between remembered contents and "immediately present" contents is dependent upon recognizing the differences between their phenomenal characters, the knowledge required to make such

[32]This is an admittedly quick gloss on a nuanced topic—the sense of "being from memory" that is attached to recollected contents is precisely what we have based our abstract concept of memory on, and thus is clearly not as insignificant as my argument might make it seem. That said, unless we are looking at remembered contents specifically as memories (such as when we say things like, "That looks familiar, but my memory is sort of fuzzy") the fact that the content is being presented *via* memory does not appear to place further justificatory demands on how we understand that content.

distinctions between modes of content presentation is not empirical.[33] It is *a priori* that sensations observed as distinct from one another must in fact be distinct, and because we (at least typically) judge that our knowledge of the present state of the world is limited to whatever is being reported by our immediate perceptual and rational faculties, then it is *a priori* that any information that we receive about the world through a means that is phenomenologically distinct from our immediate perceptual and rational faculties (*e.g.*, information supplied within a $MemoryPresents$ context) cannot be information about the present state of the world. Therefore, it is *a priori* that recollected information is not about the present state of the world, and it is thus *a priori* that if memory is veridical then the information it presents is about the past. To put this in a more straightforward way, we have an *a priori* entitlement to believe that the information presented to us within the $MemoryPresents$ context is information about what was previously the case. Thus justification is preserved through our inferences from $\exists e(e\ MemoryPresents(p))$ to $prev(p)$.

We have seen that even though we must derive our warrant for this inference from necessarily *a posteriori* understandings of the distinctions between our sensations, we do not employ empirical beliefs when inferring $prev(p)$ from $\exists e(e\ MemoryPresents(p))$, and thus it will generally be the case that we are *a priori* entitled to make this inference. It is worth noting that this inference does not necessarily enjoy *a priori* warrant in all situations, however, for it is possible that certain memorial experiences will be presented alongside a sensation of uncertainty or doubt in whether what is being presented actually corresponds to what was previously true. Suppose, for example, that I vividly remember going on a tour of a chocolate factory when I was very young, but that due to a sense of doubt accompanying my memorial experience I am no more confident that my memorial experience represents something that was previously true than I am confident that it represents a dream or an imagining that I once had. If part of my memory experience is a sense of doubt about whether the content of my recollection actually represents of a past event or previously acquired information, then I am clearly not *a priori* entitled to assume that what that content represents actually happened; rather, I will need to rely on other information (such as my beliefs about memory, whether I can produce further supporting memories, etc.) in order to be warranted in inferring that my memory is actually reporting what was previously the case. This is to say that when part of a memorial experience is a sense of doubt about the content presented by that experience, further inferences are required in order for one to be warranted in eliminating the memory trace signature operator, and thus one no longer necessarily enjoys *a priori* entitlement to infer $prev(c)$ from $\exists e(e\ MemoryPresents(c))$.[34] So long as recollected content is presented without any

[33] We can, for example, imagine a mind that has no sense perceptual faculties but which entertains itself by proving *a priori* knowable logical truths. Because we can imagine that such a mind might be able to remember the logical truths that it has already proven, and distinguish this mode of recognizing logical truths from the recognition of logical truths through proof-driven rational exercises, it appears that recognizing memory as a distinct mode of presenting information does not depend on our capacities for empirical investigation.

[34] For further discussion of this point, one might refer to Matthew McGrath's "Memory and Epistemic Conservatism" (McGrath, Matthew. 2007. "Memory and Epistemic Conservatism." *Synthese*

sense of uncertainty, however, our warrant in eliminating the memory trace signature is not defeated, such that we commonly have *a priori* entitlement to believing that the content represented within the $MemoryPresents$ context was previously true. Finally, because we have seen that at least in certain cases the inference from $prev(p)$ to p is *a priori* warranted, we see that it is possible for memory to perfectly preserve justification all the way through the memorial process—that is, if a memory trace m is sculpted with respect to a mental state representing p with degree of justification j, it is possible (depending on the content of p) for the reactivation of m to produce a mental state representing p with degree of justification j.

2.7 Memory Traces and Preservative Memory

Given the Memory Justification Principle and the fact that certain kinds of propositions (forever-after propositions, at least) can undergo memory trace signature elimination without compromising their justificatory statuses, we may now return to our considerations of Burge's claim that we employ a preservative memory faculty—a mechanism which "preserves beliefs with their justifications, but contributes no independent source of justification" (Burge, 37). Interpreting Burge's notion of preservative memory as a strong claim about the possibility of belief contents persisting in memory, our acceptance of memory traces *prima facie* rejects the possibility of preservative memory, since memory trace theory posits that only memory traces actually persist through the memorial process. Because memory traces are not themselves beliefs, but can only be associated with beliefs through causal correspondence, memory trace theory simply denies that belief contents themselves are preserved by memory. We may, however, interpret Burge's claim that preservative memory "preserves beliefs with their justifications" in a way that is compatible with memory trace theory.[35] Insofar as a reactivated memory trace can provide content that is, after *a priori* warranted tense operator elimination, identical to some of the content of the

157:1-24.), in which McGrath argues (among other things) that whenever an agent is presented with a belief from memory, "retaining the belief is rationally appropriate for [that agent]," so long as holding that belief is rational "*in light of [that agent's] current epistemic perspective*" (5, McGrath's emphasis). What McGrath appears to be noting here is that one is *a priori* entitled (*i.e.*, rational in retaining) one's memory-beliefs so long as one does not presently acknowledge the existence of any defeaters (such as a sense of uncertainty) for those memory-beliefs.

[35]There are actually two ways of understanding this claim. The most literal reading (which I will only mention in this footnote) of Burge's claim is just that there are situations in which we have *a priori* entitlement to what I have called the memory trace signature elimination, whereby we inherit a rational "right to believe" the information that memory presents to us in virtue of the fact that memory presents that information as veridical. This is the most literal interpretation of Burge because he does not attend to the matter of memory traces, and instead focuses on memorial content as what I have referred to as "the content within the $MemoryPresents$ context." In a sense, then, my argument regarding when we are *a priori* entitled to eliminate the memory trace signature operator is an argument about when it is that we employ preservative memory. Because, in the aforementioned sense, I have already defended preservative memory, I focus this section on the stronger claim of "perfectly" preservative memory, which gives greater weight to the idea that justification for memory-belief contents (not just entitlement to believe them) can be preserved through the memorial process.

mental state with respect to which the memory trace was sculpted, we can think of the ideal functioning of a memory trace as approximating the "perfect preservation"[36] of the content of the mental state with respect to which the memory trace was sculpted. Arriving at remembered content that is entirely identical to content which was represented by the mental state with respect to which a corresponding memory trace was sculpted requires the inferential elimination of a memory trace signature, however, which can be done in a justification preserving way (as we have seen in the cases of forever-after propositions). If we use this weaker interpretation of what it is for memory to be perfectly preservative, the Memory Justification Principle shows us that there is only one sort of case in which agents employ a preservative memory faculty, which may be characterized as follows:

> Perfectly preservative memory has been employed in producing a mental state s with content $sig(c)$, where sig is the memory trace signature operator iff s was produced by the activation of a memory trace that was sculpted with respect to a mental state that represented c, and it is *a priori* that $sig(c) \Rightarrow c$.

According to the Memory Justification Principle, the addition of any content other than the memory trace signature to a recollected mental state will cause a decrease in justification, and thus whenever perfectly preservative memory is employed it is necessary that the recollected mental state reflects the content of the mental state with respect to which the relevant memory trace was sculpted. Additionally, in order for the previous belief to be perfectly preserved, it is necessary that the memory trace signature can be eliminated *a priori* from the recollected content. Given that $sig(c) \Rightarrow prev(c)$, however, only mental state contents for which tense operators can be *a priori* eliminated (*e.g.*, forever-after propositions) stand as genuine candidates for being perfectly preserved by memory. Thus only recollected mental state contents which satisfy the aforementioned criteria are said to have been produced by a perfectly preservative memorial mechanism.

While the foregoing reveals that one consequence of the Memory Justification Principle is that it commits us to the claim that the kind of memory employed by human-like minds can be perfectly preservative, I believe that the relevant notion of preservative memory will not be necessarily unpalatable for philosophers whose intuitions are more closely aligned with the views of Christensen and Kornblith. We only employ a memory mechanism that perfectly preserves justification (including *a priori* warrant) when we recollect contents for which $prev(c) \iff c$ where c was fully represented by the mental state with respect to which the relevant memory trace was sculpted and c enjoyed *a priori* warrant at the time that the relevant memory trace was sculpted. Thus we avoid the " explosion of *a prioricity*" that Christensen and Kornblith are worried would be entailed by accepting a theory of preservative

[36] I will use the term "perfectly preservative memory" to signify a mechanism that (1.) provides an agent access to a mental state content that is identical to the content of the mental state with respect to which the relevant memory trace was sculpted, and (2.) connects a justificatory status to the recollected content that is equivalent to the justificatory status that was enjoyed by the original content at the time when the relevant memory trace was sculpted.

memory (14). Indeed, in most cases we would expect beliefs to lose justification during recollection, due to the empirical assumptions required for the elimination of memory trace signature operators.[37] If we take Christensen and Kornblith's concern to be tied to a suspicion that endorsing preservative memory will result in our having mental state contents that gain justification through the process of recollection, then we see that their concern is wholly unnecessary—the Memory Justification Principle caps the justificatory statuses of remembered contents at the degree of justification they enjoyed when their corresponding memory traces were formed, and thus gains in justificatory status cannot be caused by the memorial process alone.[38] Between the dictates of the Memory Justification Principle and the requirements for *a priori* warranted elimination of memory trace signatures, we have arrived at an intuitive picture of how justification is transmitted through the causal process of memory trace sculpting and activation.

2.8 Memory Traces and the SFM

Is the SFM compatible with memory trace theory? Recall that the SFM regarded agents' belief corpuses as sets of beliefs, which can be activated by relevance querying based on similarities in the various beliefs' sub-propositional constituents. Without qualification, memory trace theory is incompatible with the idea that agents have sets of stored *beliefs*; memory traces are by definition not mental state contents, and thus the literal existence of a *belief* corpus is *prima facie* rejected by memory trace theory. The qualification that patches over this apparent conflict, then, is simply that as a model the SFM must be allowed some leniency in its abstract illustrations of the mind's functions. We might think of the SFM as idealizing its representations of agents' belief structures by treating agents as though their memorial faculties always sculpt and activate memory traces with perfect precision, such that content is never gained or lost in the course of storing and accessing individual mental state contents. Similarly, we may think that the particular features of memory traces that are used to actually determine relevance roughly correspond to the sub-propositional constituents of the contents that they produce when activated. As a functional tool for talking about the belief structures of agents, the SFM certainly works with a simplistic picture of the mind, but its simplifications are unproblematic and easily translate into the more metaphysically rigorous language of memory traces. For discussions that are only concerned with the doxastic structures of agents, the SFM is sufficient, and when we are interested in the more complex project of tracing justification through an agent's memorial mechanisms the SFM provides no impediment to theorizing

[37] Take my belief that my car is currently parked in my driveway. I cannot see my car, and thus must remember that my car was last seen on my driveway in order to hold this belief. To believe that my car is still on my driveway, then, I must assume that my car has not moved unbeknownst to me, which is certainly not *a priori*, and thus my justification for believing that my car is still on my driveway will slightly suffer.

[38] In a later chapter I will consider a special sort of case in which remembered beliefs enjoy an increase in justificatory status even though the recollecting agent does nothing more than remember those contents.

about the causal histories and justificatory statuses of an agent's memory-beliefs.

This acknowledgment that the conflict between memory trace theory and the SFM is relatively superficial may not satisfy certain sensibilities, however, so I will briefly sketch an alternative *Memory Trace Model* (MTM) that captures all of the intuitions behind the SFM, but which relies on the more metaphysically rigorous language of memory traces, rather than the "deactivated beliefs" of the SFM. Let an agent's *memory trace corpus* be the set of memory traces which that agent has cognitive access to (*i.e.*, could activate, given the agent's cognitive architecture). Let an agent's *active mental state* be the sum of mental state contents that constitute that agent's present conscious experience. We can say that an agent believes a proposition p iff p is a component of the agent's active mental state or the agent has a memory trace that is disposed to produce the content p when activated.[39] Relevance between a content c and a memory trace m within an agent's memory trace corpus is defined as a binary operation that returns *true* iff the activation of m would produce a content that shares sub-propositional content with c, and returns $false$ otherwise. Relevance querying on a content c is the operation of appending to the active mental state the contents produced by activating memory traces that are relevant to c. A memory trace's centrality to a query on a content c is the probability that the memory trace will activate during a relevance query on c, and may be thought of as determined by factors similar to those which were taken to determine centrality for the SFM—frequency of the memory trace's activation, how recently the memory trace has been activated, etc. Affirmation of a mental state content may be assumed sufficient for a memory trace to be sculpted with respect to that content and added to the memory trace corpus. Lastly, the MTM, like the SFM, is neutral regarding whether there is a mechanism that deletes infrequently activated memory traces from the memory trace corpus. While this is a very brief gloss on what a memory trace model of cognition might look like, what is important is that all of the crucial operations and components of the SFM can translate easily into the language of memory traces. In a sense, the SFM is a cognitive model which assumes that memory traces always ideally preserve the contents with respect to which they are sculpted, and thus does not suffer by speaking directly to the mental state contents that would be produced by the activation of those idealized memory traces. The conflict between the SFM and memory trace theory is merely a superficial conflict based on their relative levels of abstraction, and we may thus maintain the SFM as a useful means for modeling the fragmented belief structures of agents.

We may also note that the notion of memory traces supports the SFM's stance that agents are only committed to the consequences of closing the conjunction of their *active* beliefs under unrestricted classical implication. According to memory trace theory, the term "deactivated beliefs" is only used as a shorthand means for expressing for something along the lines of "memory traces sculpted with respect to mental states that represented particular belief-contents and that are disposed under ideal

[39]We may also wish to add the qualification that if p is a timeless proposition and the agent has a memory trace that is disposed to produce the content $prev(p)$ then the agent believes p. I will discuss this matter in greater detail in the following section.

circumstances to induce mental states that represent those same particular belief-contents." Within memory trace theory we thus have no reason to treat the agents we consider as though there is any logical interaction between their "deactivated beliefs," and we may see that memory trace theory supports the SFM's claim that we cannot commit agents to the logical closures of their complete belief corpuses. Furthermore, although the SFM does commit the agents it models to the explicit contents of their deactivated beliefs, this seems acceptable so long as we accept the SFM as an idealized model in which memory traces will always produce contents that are identical to the contents of the mental states with respect to which they were sculpted. Therefore, in addition to being compatible with the SFM, memory trace theory reinforces the argument that the SFM is a highly plausible approach to modeling actual agents' doxastic structures.

2.9 Latent Belief Attribution

These considerations raise a further question, however, of what the necessary and sufficient conditions are for the correctness of latent belief ascriptions[40] when we are working with the notion that "deactivated beliefs" are actually non-contentful memory traces. At first blush, we may be tempted to simply say that an agent has a latent belief that p iff the agent has a memory trace m that was sculpted with respect to a belief state representing p. There are certain cases in which this is intuitively insufficient. Consider the possibility that I have a memory trace m that was sculpted with respect to a mental state that represented the belief "There are eight planets in our solar system." Due to a malfunction in my cognitive architecture, however, I am unable to reproduce this mental state by activating m—perhaps m produces a mental state that represents a bunch of swirling colors, or perhaps m's activation contributes nothing to my present mental state. m has not changed in any way, but due to some odd feature of my mind, I am unable to reproduce the belief with respect to which m was sculpted. Assuming that I cannot reproduce the belief content "There are eight planets in our solar system" without the proper activation of m, then it seems intuitive that I should not be said to have a latent belief that there are eight planets in the solar system. Prior to undergoing some adjustment of my cognitive machinery or re-learning the number of planets in our solar system, I simply cannot stand in a memory-generated doxastic relation to the propositional content "There are eight planets in our solar system." We thus see that an agent's possession of a memory trace that was sculpted with respect to a belief state representing p is insufficient for our ascription of a latent belief that p to that agent.

We may also note that the foregoing account is insensitive to the role of tense in latent belief ascriptions, insofar as it delivers the result that an agent who sculpted a

[40]By "latent belief ascriptions" I mean attributing particular beliefs to agents when those beliefs are not active in thought or behavior-guidance. When an agent is giving an art history lecture, for instance, we typically say that the agent holds beliefs about things like who the President is (if she's American) and how many planets there are in the solar system, even though she is almost certainly not actively considering these beliefs.

memory trace with respect to a belief state representing "I am at the beach" would be said to have a latent belief that he *is* at the beach, rather than that he *was* at the beach, which clearly does not conform to our intuitions. Seeing that we also frequently ascribe present-tense latent beliefs to agents, we cannot simply say that all latent beliefs have a tense operator attached to them. We will thus require that our account of latent belief does not demand that all latent beliefs are past-tense or that all latent beliefs are present-tense. To help us see how we might develop some such account, we might recall that the way we accounted for present-tense recollections was by an appeal to the fact that in at least some cases $prev(p) \Rightarrow p$. Assuming that the propositions p for which $prev(p)$ implies p will also, generally speaking, be the ps for which we intuit that agents have present-tense latent beliefs that p, and given that making the inference from $prev(p)$ to p requires that the inferring agent believes $prev(p)$, then it appears that when we ascribe a latent belief in a present-tense p to an agent, this ascription is in virtue of the fact that the inclusion of $prev$ in the recollected content is an unnecessary technicality. Insofar as we have to admit that agents who latently believe present-tense propositions may also be taken to believe the past-tensed versions of those propositions, we appear to be saying that agents latently believe the consequences of whatever they are capable of being presented by memory. For these reasons, let us use the logical consequence operator cn, which takes as argument a (possibly empty) set of propositions and returns the set of propositions that are logical consequences of the argument, and define latent belief in the following way:

> Let c be a propositional content and let sig be the memory trace signature operator. We say that an agent A latently believes each non-experiential[41] element in $cn(\{sig(c)\})$ iff there exists some memory trace m that was sculpted with respect to a belief state that represented p and A's cognitive architecture is such that m could be activated to produce a mental state representing $sig(c)$.

Plainly speaking, this says that if an agent has a memory trace that could produce a belief state with content identical (adding considerations of the memory trace signature) to the belief state which was represented by the mental state with respect to which the memory trace was sculpted, then that agent has by definition a latent belief in each of the logical consequences of that content. While this commits agents to latent beliefs in some admittedly strange propositions in virtue of the nature of logical implication (*e.g.*, bizarre disjunctions, even-numbered negations, etc.), this does not seem terribly troubling. While it seems somewhat unusual to say that I believe the proposition "Abe Lincoln is not not a former P.o.t.U.S., or all birds are ravens" in

[41]By this restriction on "experiential" contents I mean to preclude contents that posit the existence of particular sense experiences from being attributable to agents as latent beliefs. We do not wish to say that an agent who has a memory trace that is disposed to produce the content $\exists e(e\ MemoryPresents$(My name is Alex)) has a latent belief that he is experiencing having memory present to him that his name is Alex, although we do wish to say that such an agent has a latent belief that his name is Alex. It is this distinction that I intend for my use of "non-experiential" to capture.

virtue of my prior affirmation that Lincoln is a former P.o.t.U.S., as a *latent* belief ascription there is nothing wrong with saying this. In fact, I imagine that if most agents familiar with propositional logic and American history were asked, "Do you believe the statement 'That Abe Lincoln is not not a former P.o.t.U.s., or all birds are ravens?'" they would affirm that this is a proposition that they believe, even if it took them a moment to parse the meaning of the statement. The effects of using the consequence operator *cn* in our account of latent belief attribution are thus not as troublesome as they might appear at first.

There is, however, at least one even stranger consequence of this account of latent belief—it commits me to latently believing any belief contents which I have sculpted preservative and activatable memory traces with respect to, even if I have since realized that those contents are false. Suppose, for example, that when I was young my friend Shaun told me that his parents were cyborgs, and that I sculpted a memory trace with respect to the mental state representing "Shaun's parents are cyborgs." I am currently able to recognize that Shaun's story about his parents was obviously made up, but so long as I have the relevant memory trace I am committed to latently believing $cn(sig(\text{Shaun's parents are cyborgs}))$. Supposing that the memory trace signature analyzes to *prev*, I am thus committed to a latent belief that $prev(\text{Shaun's parents are cyborgs})$, as well as this belief's logical consequences. While this may seem somewhat strange at first, I do not believe that this consequence is unacceptable, or even genuinely counterintuitive. If I possess a memory trace that is capable of producing the content $prev(\text{Shaun's parents are cyborgs})$, then there is a non-zero chance that if I were asked whether I believe Shaun's parents are cyborgs I would respond in the affirmative. I would almost certainly recognize after recalling my other cyborg-relevant beliefs that my response had been ridiculous, but so long as I have a cognitive disposition to affirm that Shaun's parents are cyborgs it seems incorrect to say that I do not latently believe that Shaun's parents are cyborgs (of course, I also latently believe Shaun's parents are not cyborgs). While we like to think that agents are capable of erasing their prior epistemic shortcomings, this seems to be an idealization based on what we would like to be the case rather than how our minds actually function, and thus we should not be shocked to find ourselves latently committed to beliefs that we have condemned.[42] The aforementioned consequences of the memory trace account of latent belief attribution may be found counterintuitive at first glance, but upon closer inspection we may realize that these consequences do not actually run contrary to our intuitions about belief attribution.[43] At the very least,

[42]To use a less cartoonish example, we can imagine an agent who was raised in an environment where she was taught that homosexuals are evil, and who formed some such prejudicial belief on the basis of others' misinformation. As the agent investigates the issue for herself and meets gay people who are not morally corrupt, she may wish to change her beliefs about homosexuals. Unless we wish to say that prejudice can just be willed away at an instant, however, it is intuitive to maintain that the agent will continue to harbor the latent belief "Homosexuals are evil" for some amount of time after her change in attitude. While it may still seem odd that we have to say of this now open-minded agent that she has a latent homophobic belief, this simply warns us not to think of singular latent beliefs as being representative of agents' full doxastic landscapes.

[43]It is worth remarking that it is conceivable that certain mental state changes could effect changes in relevant memory traces. When I first realize that Shaun's parents are not actually cyborgs, for

my discussion does not reveal any *prima facie* reasons to reject this account of latent belief. For these reasons I maintain that this memory trace account of latent belief attribution is correct.

We may thus note that memory trace theory supports the adequacy of the SFM in two ways. First, we saw that memory trace theory supports the theory that agents should not be committed to the elements in the logical closure of the set of their deactivated beliefs, since memory traces are not themselves contentful and thus should not be thought of as logically interacting with one another. In this section we have seen that memory trace theory provides us an account of latent belief that allows us to think of agents as having beliefs in the contents of what their memory traces would ideally produce upon activation, and thus the latent beliefs of an agent may be seen as analogous to the belief corpus of that agent. While the SFM certainly stands as a simplified account of the cognitive mechanisms developed by memory trace theory, there is nothing in memory trace theory that is incompatible or substantially in conflict with the SFM, and thus we may maintain that the SFM accurately models the doxastic organization of human-like minds.

2.10 Concluding Remarks

We have seen in this chapter that the Memory Justification Principle and the rules of memory trace signature elimination allow for actual agents to employ a preservative memory faculty without causing an explosion of *a prioricity*, thus protecting a weaker interpretation of Burge's theory of preservative memory while also soothing Christensen and Kornblith's concerns about the essentially reconstructive nature of memory. This is allowed by the introduction of the notion of memory traces as things that are not mental state contents but are somehow sculpted so as to induce mental state contents that are distinctly recognizable as corresponding to previously entertained contents. Further, it was found that the notion of memory traces conflicts superficially with the SFM, the SFM may be understood as a simplification of the memory trace based MTM of cognition, and thus survives as an accessible idealization of the way in which actual agents employ fragmented belief structures to make up for our limited cognitive capabilities.

example, it is conceivable that the changes taking place in my mental state during this realization could cause the "Shaun's parents are cyborgs" memory trace to be altered in such a way that it no longer qualifies me for latently believing that Shaun's parents are cyborgs—the memory trace that activated the cyborg-parents belief might be "re-sculpted" in a way that signifies that its contents have been debunked, or something along those lines. While such possibilities are merely speculative, it is worth noting that counterintuitive cases of latent belief possession like the cyborg-parents case may be rare occurrences.

Chapter 3

Normative Defeaters and an Attack on the Preservation View of Memory

3.1 The Preservation View of Memory

In "Memory as a Generative Epistemic Source," Jennifer Lackey offers an argument against what she calls the *Preservation View of Memory* (hereafter PVM).[1] The PVM is the thesis that a subject S stands in some epistemic relation (*e.g.*, knowing, believing, having a belief with a particular justificatory status) to a proposition p at time t_2 because of memory alone only if (1.) S stood in the same relation or a stronger relation of the same kind[2] to p at an earlier time t_1, and (2.) S stood in that relation to p at t_1 for a reason other than memory (637). The essential point of the PVM is that it *limits* memory to being preservative of epistemic status; if I am to know that p in virtue of memory alone, then at least one of my prior epistemic relations to p must have been knowledge of p based on something other than memory. We may note that the PVM is a necessary consequence of the Memory Justification Principle that I have endorsed, since the Memory Justification Principle claims the justificatory status of a remembered belief is calculated by multiplying the original belief's degree of justification with the proportion of the propositional content of the remembered belief that is also represented by the original content (along with appropriate considerations of memory signature content). Because this proportion cannot be greater than one, the Memory Justification Principle necessarily caps the justificatory statuses of memory-beliefs at the degree of justification enjoyed by their trace-sculpting counterparts, and thus entails the PVM.

---

[1] Lackey, Jennifer. 2005. "Memory as a Generative Epistemic Source." *Philosophy and Phenomenological Research* 70: 636-658.

[2] The first condition of the PVM is also satisfied by t_2 epistemic relations that were constitutive of t_1 epistemic relations. If belief is a necessary constituent of knowledge, for example, then (1.) will be satisfied in cases where S knew p at t_1 and believed p at t_2. Similarly, if an agent believed p with a degree of justification j_1 at t_1, and remembered p with a degree of justification j_2 at t_2, (1.) will be satisfied for all $j_2 \leq j_1$.

Suppose, for example, that at time t_1 I hold the belief that Abraham Lincoln's face is one of the faces depicted on Mount Rushmore, but that I cannot be said to know this fact (perhaps my only evidence that Lincoln is on Mount Rushmore is that questionably reliable sources have told me so). At a later time t_2, according to the PVM, it is impossible for me to know that Lincoln is on Mount Rushmore in virtue of my remembering that Lincoln is on Mount Rushmore. Without further evidence (*e.g.*, comparing a Google image search of "Mount Rushmore" with the face on my lucky penny) my epistemic status regarding Lincoln's being depicted on Mount Rushmore cannot be upgraded. This is the intuition that the PVM wishes to capture.

Another way of approaching the PVM is to note that the analogous intuition that holding a belief does not, in and of itself, establish knowledge. My belief that Batman is the best superhero does not in and of itself establish my knowledge that Batman is the best superhero. Similarly, the PVM asserts that a subject's memory can by itself only establish an epistemic status regarding some proposition that is weaker than or equal in strength to what the subject's epistemic status with regard to that proposition was on the occasion of the memory's formation. According to the PVM, my memory at t_1 that Batman is the best superhero cannot by itself ground my having knowledge at t_2 that Batman is the best superhero. Despite the PVM's intuitive features, Lackey argues that it is false. Her argument is that even though the PVM accurately captures the limits of memory's epistemic authority in many cases, it is possible for memory to perform a "generative" role—that is, there are cases where "even though memory did not generate the belief in question, it generated *the epistemic status* of the belief in question" (649, Lackey's emphasis). If Lackey is correct, then there are cases where condition (1.) of the PVM fails, and the PVM's attempt at describing the epistemic statuses of memory-beliefs falls short of universal success.

3.2 Normative Defeaters

As is the case for many initially counterintuitive arguments, Lackey's claim may be a hard pill for some to swallow. It will therefore be useful to provide a brief overview of what Lackey presupposes before presenting her argument itself.

The most challenging supposition that Lackey's argument relies upon is the existence of external normative defeaters, where a normative defeater for a belief that p is defined as "a proposition D that S ought to believe to be true, yet indicates that S's belief that p is either false or unreliably formed or sustained" (638).[3] For an illustration of a normative defeater, one might consider the following case: It is the night of a mid-term senate election, and it is my job to add up the vote tallies coming in from the different polling centers in my county. Furthermore, suppose that my

[3]In addition to normative defeaters, there are also doxastic defeaters. A doxasitic defeater is defined by Lackey as "a proposition D that is believed by S to be true, yet indicates that S's belief that p is either false or unreliably formed or sustained" (638). I take these as epistemically commonplace, and for this reason will not give this species of defeater the sort of attention that the more controversial notion of normative defeaters requires.

county is the largest and most liberal region in a predominantly conservative state, such that I am justified in believing, as every reputable news source has predicted, that in order for the Democrat to even have a shot at winning the race she has to carry 55 percent of my county's votes. While I tally the votes, I am in a room with a computer, a television, and a radio, as well as a window facing a packed bar. To avoid distraction, however, I turn off all the electronic devices, close the windows, and draw the blinds. Being a meticulous accountant, I work slowly. Time passes, and as all of my colleagues in the state's other polling districts submit in their final tallies every major local news source issues reports that the Democratic candidate only needs 48 percent of my county's vote to win, due to unprecedented low turnout in the state's most conservative areas. If I were to turn on any of the electronic devices around me, or even just open my window, I would acquire the belief that the Democratic candidate needs 48 percent of my county's vote to win. In this situation, given my strong interest in the outcome of the election and how trivial of a task it would be for me to acquire the belief that the Democratic candidate only needs 48 percent of my county's vote to win the race, I have a normative defeater for my prior belief that the Democrat needs at least 55 percent of the vote.[4] I am therefore unjustified in forming any epistemic relation to beliefs that are based on my normatively defeated belief that the Democratic candidate needs at least 55 percent of my county's vote. At the end of my calculations, however, I see that the Democratic candidate ended with 51 percent of the vote in my county and believe that the Republican candidate won.

Despite the fact that my belief that the Republican won was logically inferred using only concrete evidence and my previously justified belief that the Democrat needed at least 55 percent of my county's vote to win, it is unjustified. My belief that the Republican won, however, is not defeated in virtue of being false (false beliefs are not necessarily defeated[5]) or in virtue of using a poor method of inference (the logical formula I applied in order to arrive at the conclusion that the Republican won was internally airtight). Instead, it is defeated in virtue of there being a proposition that I *ought* to believe which defeats the belief that necessarily grounds my further inferred belief that the Republican won. This proposition that I ought to believe is capable of defeating my previously justified belief in virtue of the fact that the evidence for the defeater is so readily accessible that I am essentially committing gross epistemic negligence by forming a belief without considering it. In other words, given my present context I cannot justifiably form a belief regarding this subject without having accessed the available evidence, so by merely thinking that I have justifiably formed a belief within this context, my belief must be held to the same

[4]This situation may not appeal to everybody's epistemic sensibilities regarding sufficient conditions for the existence of a normative defeater, but one can easily add on further situational details to reach the point where it is intuitive that my belief has been defeated (*e.g.*, I am receiving countless texts from Democratic friends asking if I have seen the poll numbers but I am uncharacteristically ignoring my phone, etc.).

[5]Ancient peoples who believed that the Earth was flat held a false but undefeated belief; this belief was undefeated in virtue of the fact that those people had no reason to believe otherwise. Such a belief today, on the other hand, would be defeated.

epistemic standards governing an agent who has accessed the defeating evidence. This being the case, any belief that I form contrary to the extent normative defeater is necessarily defeated, even though I have not cognitively interacted with the defeating proposition itself.

The essential characteristic of normative defeaters is that they are operative in cases involving agents who have not satisfied the context-dependent criteria necessary for formulating certain kinds of beliefs within their contexts. Such unsatisfactory performances do not signify the shortcomings of agents *qua* agents who can take bodies of evidence and reliably assess their implications (one might call agents who use these abilities epistemic *processors*), but *qua* agents who are capable of determining when the evidence that they have at hand can be effectively processed so as to produce beliefs corresponding to the world (epistemic *investigators*). While Lackey's commitment to normative defeaters may strike some as initially unpalatable, it does not contradict standard epistemic sensibilities, and for this reason we will do well to grant the existence of normative defeaters for the time being, if only for the sake of internally analyzing Lackey's argument against the PVM. With this in mind, I will now present a Lackey-style case against the PVM.

3.3 A Lackey-Style Case Against the Preservation View of Memory

Lackey provides several cases in defense of her counter-PVM position. The structure of these cases begins with an unremarkable epistemic agent, let us call him S, who stands in some low-level epistemic relation to a proposition p at time t_1 due to the presence of a normative defeater D. S's relation to p at t_1 is due to a source other than memory, so the second condition of the PVM will be satisfied in the event that at a later time S uses his memory to reestablish an epistemic relation to p. S does, in fact, use memory to reestablish an epistemic relation to p at a later time t_2, and, furthermore, does this without acquiring any new evidence that would weigh in on his epistemic relation to p. When S consults his memory, however, the epistemic relation connecting him and p at time t_2 ends up being stronger than the relation that he stood in with respect to p at t_1 in virtue of the disappearance of D. In this case, the agent forms a belief based solely on memory that enjoys a higher degree of justification than the memory-belief's original counterpart, thus violating the first condition of the PVM. The common conclusion that we are meant to intuit from Lackey's cases is that there are possible situations where memory alone allows an agent to upgrade his epistemic status relative to some proposition, and that in virtue of this being the case, the PVM fails. I will now present and assess a Lackey-style case of generative memory.

Recall the basic setup of the vote counting case from the previous section. I have counted the votes for my county, seen that the Democrat got 51 percent, and immediately form a belief that the Republican has won at time t_1. In virtue of the fact that I have an epistemic obligation to believe that the Democrat only needed 48 percent

of the vote to win, my belief is defeated by a normative defeater. Having finished my work and feeling exhausted, I report the count and go directly home, neglecting to obtain further information about the election before I fall asleep. Suppose that while I am sleeping, the Republican candidate demands a recount, and in the course of the recount it is discovered that the reports of record-low turnout were actually the result of miscalculations caused by the hasty accounting work done by my colleagues in other districts. Thanks to my meticulous and transparent accounting, the votes for my county can be reassessed by other election officials without calling me back in, so I sleep through the recount demands of both parties. In the course of the recounts, it turns out that the Democrat actually needed 55 percent of the vote in my county in order to win the race, and, because she only received 51 percent of the vote, the Republican actually won the race. In virtue of this new information, my normative defeater has been defeated, and thus can no longer exercise any authority over the justificatory status of my beliefs.

At time t_2, when I wake up in the morning, there is no reason for me to believe that the Democrat only needed 48 percent of the vote; all of my readily accessible evidence sources would tell me that the Democrat needed at least 55 percent of the vote in my county to win the race. This is to say that I am once again justified in holding beliefs based on a requisite belief that the Democrat needed 55 percent of the vote in my county to win, since this belief is no longer defeated. Therefore, when I use memory alone at t_2 to reestablish an epistemic relation with the proposition "the Republican won," this epistemic relation will be stronger than its original counterpart was at t_1, when my belief was subject to the effects of the normative defeater. Furthermore, it is not essential that my belief at t_2 achieve a high degree of justification. What matters is simply that my belief that the Republican won the election has had its justificatory status upgraded. So long as this is the case, memory, in virtue of generating a belief that the Republican won the election which enjoys a high degree of justification at t_2, produces a belief with an epistemic status higher than the content-identical belief that I formed at time t_1 *via* non-memorial faculties. This case reveals that if we accept the existence of normative defeaters, then in virtue of recollection alone one is able to form memory-beliefs whose epistemic statuses are superior to the epistemic statuses of their original counterparts. This directly contradicts condition (1.) of the PVM, and thus the PVM fails to accurately characterize the transmission of epistemic status in all cases of recollection.

Lackey's considerations of normative defeaters thus generate serious problems for proponents of PVM. Given that the Memory Justification Principle entails the PVM, this is deeply troubling for the account of the transmission of justification through the memorial process that I previously developed. One method of resisting Lackey-style cases that is readily available to proponents of the PVM is to reject the existence of epistemic obligations and, by extension, the normative defeaters that make Lackey-style cases possible. I will now turn to investigate whether this is a viable means to resist Lackey's argument against the PVM.

3.4 Epistemic Investigation

The crux that Lackey's argument relies upon is the capacity for an agent's unfulfilled epistemic obligations to ground normative defeaters, which can then exert undetected influences on the epistemic statuses of certain of the agent's beliefs. Lackey, explaining the essential nature of defeaters, writes, "The underlying thought here is that certain kinds of counterbeliefs and counterevidence contribute epistemically unacceptable *irrationality* to doxastic systems and, accordingly, that justification and knowledge can be defeated or undermined by their presence" (639, Lackey's emphasis). While this provides an effective overview of how defeaters will have to function in Lackey's argument, an account of the process by which epistemic obligation can defeat an agent's beliefs is unspecified. Given that Lackey's cases require us to find the connection between epistemic obligation and normative defeat viable, it will be important to have at hand a method through which one may connect epistemic obligation to the defeat or undermining of an agent's beliefs in appropriate situations.

As I mentioned before, two components of epistemic agency are epistemic processing and epistemic investigating, and the primary fault of agents in Lackey-style cases is that they neglect their obligations as epistemic investigators. Outside of Lackey-style situations, however, we primarily judge agents based their activities as epistemic processors—that is, on whether they can reliably form rational judgements based on their given evidence bases—and rarely invoke the notion of standard investigatory procedures, so I will now attempt to nail down exactly what I mean in my considerations of agents *qua* epistemic investigators. To examine the distinguishing features of epistemic investigation, I will make use of the notion of an "epistemically responsible agency," as developed by Hilary Kornblith.[6] "An epistemically responsible agent," according to Kornblith, "desires to have true beliefs, and thus desires to have his beliefs produced by processes which lead to true beliefs; *his actions are guided by these desires*" (Kornblith 34, emphasis added). While Kornblith intends for his notion of epistemically responsible agency to encompass all aspects of epistemic agency, for our present purposes the scope of his claim may easily be narrowed so as to emphasize the investigative component of epistemic agency: insofar as an agent's conscious awareness of a certain state of affairs is based on a disposition or desire to understand the world to the best of her abilities, that awareness constitutes epistemic investigation. In other words, insofar as an agent can actively determine her exposure to certain states of affairs, her dispositions and desires for true belief formation may be understood as guiding her in establishing sufficient evidence to make and maintain well-formed beliefs. We thus see that epistemically responsible investigatory behaviors are essential for agents to exercise genuine responsible epistemic agency.

Suppose, for illustration, that I am relaxing with some friends in my living room on a Friday night, and one of my friends expresses that he particularly liked the beer that he just finished and asks me if there are any more in the refrigerator. I cannot remember, so I get up and go to the refrigerator to check. I walk through the doorway

[6]Kornblith, Hilary. 1983. "Justified Belief and Epistemically Responsible Action." *The Philosophical Review* 92: 33-48.

into the kitchen and open the refrigerator door. While looking to see if we have any more of the particular kind of beer, I notice a container of leftover fried rice from a few nights ago, and think to myself, "I'm sure that's still good to eat." I then see a pack of the beers my friend liked, grab it, and rejoin the group. This case explicitly mentions a number of investigatory acts. Several such acts are my awareness of the fried rice in the refrigerator, my remembering when I first acquired the rice, and my awareness that there were more beers of the desired kind. Each of these particular awarenesses constitutes an act of epistemic investigation; even though my primary focus was simply to determine whether or not there were more beers of the kind that my friend had, my disposition to form true beliefs about the world played an operative role in the other states of affairs that I noticed, such as noticing the rice and then remembering how old it was in order to know whether it was still good. I had no explicit interest in believing anything about the rice, but my broad interest in knowing about the world allowed my visual awareness of the rice (as well as my recollection of its origins) to constitute epistemic investigation. Some of my investigatory acts did not constitute acts of epistemic investigation, however, such as my awareness of the kitchen doorway. That is to say, I did not at any time form a belief about there being a doorway, even though my behavior required a "this is how to get to the kitchen" sort of awareness, whereby I made sense of a relevant state of affairs in a perfectly non-epistemic way. It is this sort of unreflective awareness that is paradigmatic of non-epistemic investigation, and I agree with Kornblith that there is an important distinction to be made between epistemic investigations and investigations which are not guided by our dispositions or desires to form true beliefs.

We have seen that in order for an investigation to qualify as an epistemic investigation, the investigating agent need only have her evidence-collecting behaviors guided by particular internal mechanisms—namely, those mechanisms which compel us to acquire information about the world for the purpose of forming true beliefs. While these belief-disposing mechanisms are inscrutable and vaguely defined components of the mind, they are clearly part of our psyche and distinct from the investigatory mechanisms that are not oriented towards forming true beliefs. The fact that our understanding of the mechanisms that distinguish epistemic investigation from mere investigation is vague does not make our use of the terms troubled, however, for our concern is not with the mechanisms themselves. All that we need in order to proceed with the notion that agents can be more or less responsible epistemic investigators is to allow that we have certain mental mechanisms which dispose us to form beliefs and that these mechanisms are distinguishable from certain of our other investigative mechanisms, such that when we apply these belief-disposing investigative mechanisms through the course of our investigatory practices we may said to be behaving *qua* epistemic investigators.

The significance of accepting that we can behave as epistemic investigators is that if certain acts of investigation may be said to be exercises of epistemic agency, then these actions become liable to contribute to the justificatory statuses of the beliefs that one forms based on one's gathered evidence. Kornblith, after constructing the notion of an epistemically responsible agent, goes on to write, "Being justified requires more than simply reasoning properly; it requires that one gather evidence properly

as well" (35, emphasis added). Kornblith not only affirms the importance of the investigatory component of epistemic agency, but also indicates that there are norms involved with this notion that can play an influential role in determining the degree to which of an agent's beliefs are justified. These are the same norms that Lackey-style cases use to ground their appeals to the influence of normative defeaters. With this in mind, I will now turn to investigate how the norms surrounding our acts of epistemic investigation (*i.e.*, our obligations as epistemic investigators) help determine the justificatory statuses of our beliefs.

3.5 Epistemic Obligation

Epistemic obligation, like moral obligation, prescribes the proper behavior of an agent. The obligations of an epistemic investigator, in particular, concern the norms for examining evidence sources that might contribute to the formation of certain beliefs. In practice, these obligations are relatively lax, so the situations in which we are prone to consider their role are frequently situations where some remarkable negligence has occurred on the part of an agent. Suppose, for example, that I am at a modern art gallery and that I have come across a room containing several works by a rising star whom I have never heard about but whose work is well respected and commonly discussed within the art community. When I ask one of the museum interns some general questions about the artist, he may find it disappointing that laypeople have not heard of this influential new figure, but we would not typically say that he would be reasonable to maintain that I ought to have known about the artist. There is no general expectation that a properly functioning agent in my situation would have accessed any body of information concerning this artist, and thus my ignorance does not violate my obligations as an epistemic agent. On the other hand, suppose that one of the artist's most famous pieces is a "found-sculpture" work which incorporates a comfortable-looking couch, and that the museum has placed several large placards around the couch that read, "Part of Exhibit – Not for Sitting!" with diagrams of seated figures with 'X's marked through them, so as to deter patrons from interacting with the artwork. When I sit down on the couch to rest, the museum intern will most certainly maintain that I *ought* to have known that the couch was not intended to be used as furniture, even though I simply did not notice the signs. Regardless of my ignorance of the art world, I clearly behaved in violation of one of my basic obligations as a belief-forming agent when I neglected to notice that the couch was not intended for sitting. Epistemic agents may have relaxed investigative obligations through the course of their standard operations, but these obligations are nevertheless present and serve to determine whether a person is properly exercising his epistemic capabilities. Seeing that the existence of investigative epistemic obligations is not immediately objectionable, I will shift my focus back towards an examination the mechanisms working in Lackey's argument against the PVM, and, in particular, explaining how it is that acting contrary to one's basic epistemic obligations can cause one's beliefs to be unjustified.

We may note that when an agent performs an epistemic action that contradicts

his obligations, he is behaving in a way that tends to introduce irrationality into his doxastic system.[7] This follows from what it is for something to be an epistemic obligation, since epistemic obligations are the standards to which an agent must comply if he is to consistently be in a position from which he may behave in an epistemically permissible (*i.e.*, rational) manner, such that noncompliance prevents agents from being able to regularly generate rational beliefs. Given that leaving one's epistemic obligations unfulfilled tends to introduce irrationality into a doxasitic system, any behaviors that contradict our epistemic obligations must be deemed unjustified—they simply ought not be performed. There is, therefore, a pertinent question stemming from the notion that agents epistemic obligations have epistemic obligations which will be worth considering in light of our present investigation: To what extent do an agent's epistemic obligations bear on the epistemic statuses of his beliefs?

To understand how epistemic obligation can influence the justificatory status of one's beliefs, we must determine a principle bridging how an agent behaves in forming her beliefs to the justificatory statuses of the beliefs that her behaviors produce. The principle that I briefly advertised in section 2 maintains that the analysis of beliefs produced by unjustified acts of belief formation must be performed as though the perpetrator had not behaved contrary to her epistemic obligations. Therefore, in cases where an agent has unfulfilled investigative obligations, her beliefs enjoy the same degree of justification that they would have enjoyed if the agent had fulfilled her epistemic obligations, but, despite having investigated the obligatory evidence sources, still formed the same beliefs that she actually did for the same reasons that she actually did (*i.e.*, with disregard to the obligatory evidence sources).[8] To phrase this more formally:

> *As-If Principle:*
> If an agent A has an epistemic obligation O then for any belief b that A holds in virtue of some method of reasoning over a set of evidence, $r(E)$, the justificatory status of b cannot exceed the justificatory status that b would enjoy if A had met O and had formed b in virtue of $r(E)$.

I refer to this principle as the *As-If Principle* because its key feature is that it treats the beliefs of agents with unfulfilled obligations as though those agents had formed the same beliefs through an identical process despite having formed the beliefs which would have coincided with the fulfillment of their epistemic obligations—in a sense, the As-If Principle treats agents as though they are ignoring certain of their active beliefs through the course of their belief formation activities. The utility of the As-If Principle is that it serves to connect our understandings of agents' epistemic

[7] Borrowing from Lackey's phrase book (639).

[8] Even though agents rarely, if ever, know the full extent of the evidence and reasoning operative in the production of any belief, it is not inconceivable that an agent could fully (consciously and subconsciously) disregard some acknowledged evidence source. I therefore assume that there are no *prima facie* objections to this principle.

obligations with the epistemic statuses of their beliefs. Whether or not we actually adopt this principle hinges on its compatibility with our intuitions about justification. In the following section, I will work to show that the As-If Principle conforms to these intuitions and that there is thus no *prima facie* reason to reject it. Assuming that the As-If Principle survives under scrutiny, we will have secured a method for assessing the impacts of epistemic obligations on our belief structures, and, by extension, we will possess a means for understanding the effects of normative defeaters on our beliefs.

3.6 Example Applications of the As-If Principle

Does the As-If Principle do what we want it to in cases of unfulfilled epistemic obligations as well as cases of fulfilled epistemic obligations? To answer this question, I will begin by considering two formal cases as examples. For both cases, suppose that an agent A has an epistemic obligation O that would be satisfied on the condition that evidence items e_1 and e_2 are constituents of A's total set of evidence, E_A. That is, $e_1, e_2 \in E_A$ satisfies O for A. Further, suppose that there is a proposition $e_0 = \neg e_1$, as well as a proposition e_3, neither of which A is obligated to believe. As a notational aside, I will represent the ordered pairing of a belief in a proposition P and the justificatory strength of that belief j with the tuple $< P, j >$, where $j \in [0, 1]$ and may be understood as a simplified numerical representation of the belief's degree of justification.[9] Finally, suppose that A's standard methods of rational analysis, represented as the function r_A over a set of evidence, are such that for propositions p and q:

1. $r_A(E_A) \implies < p, 0.7 > \textit{and} < q, 0 >$, for E_A such that $e_1, e_2 \in E_A$ and $e_0, e_3 \notin E_A$
   (*i.e.*, only the obligated evidence items are in A's total relevant evidence set.)

2. $r_A(E_A) \implies < p, 0 > \textit{and} < q, 0.7 >$, for E_A such that $e_0 \in E_A$ and $e_1, e_2, e_3 \notin E_A$
   (*i.e.*, only the item that contradicts e_1 is in A's total relevant evidence set.)

3. $r_A(E_A) \implies < p, 0.4 > \textit{and} < q, 0 >$, for E_A such that $e_2 \in E_A$ and $e_0, e_1, e_3 \notin E_A$
   (*i.e.*, only one of the obligated evidence items—namely, the one that does not contradict e_0—is in A's total relevant evidence set.)

4. $r_A(E_A) \implies < p, 0.9 > \textit{and} < q, 0 >$, for E_A such that $e_1, e_2, e_3 \in E_A$ and $e_0 \notin E_A$

[9] In this way, certain decimal values might be thought of as thresholds for certain degrees of justification. For example $x \leq 0.5$ may correspond to an unjustified belief and $0.5 < x$ may be correspond to a justified belief.

(*i.e.*, only the item that contradicts e_1 is excluded from A's total relevant evidence set.)

With all this in place, my goal is to illustrate that the As-If Principle undermines the justificatory status of one's belief structure when what one is obligated to believe and what one actually does believe comes into conflict (this will be the content of the first example), but does not get in the way of beliefs attaining justificatory statuses higher than those that would result from holding *only* the baseline beliefs necessary for the satisfaction of one's epistemic obligations (the second example).

For our first example, assume that agent A's actual evidence base E_A is such that $e_0 \in E_A$ and $e_1, e_2, e_3 \notin E_A$. We know from item 2 in the above list of implications that this case yields the result that A is justified in believing p to degree 0 and justified in believing q to degree 0.7, prior to any interference from A's epistemic obligations. There is interference, however, given our supposition that A is subject to an epistemic obligation that is only satisfied once $e_1, e_2 \in E_A$. According to the As-If Principle, the justificatory status of A's beliefs will be determined by treating A as though e_0, e_1, and e_2 are all members of his evidence base E_A. We may recall that $e_0 = \neg e_1$, which is to say that e_0 and e_1 are contradictory. Let us assume that whenever an agent explicitly believes contradictory propositions,[10] the degree of justification enjoyed by each of the contradictory beliefs is set to 0, and we treat the agent's other beliefs as though they derive none of their justification from the contradictory beliefs.[11] I will discuss this claim in greater detail in section 7, but assuming that it holds we see that because e_0 and e_1 are contradictory, the only member of A's relevant belief set that influences the justificatory statuses of p and q is e_2—this is to say that we end up treating A as though his evidence base is that described in item 3 above. Recall from item 3 that we have assumed that when $e_2 \in E_A$ and $e_0, e_1, e_3 \notin E_A$, $r_A(E_A)$ results in A being justified in believing p to degree 0.4 and justified in believing q to degree 0. In virtue of the As-If Principle, we see that the justificatory statuses of A's beliefs cannot *exceed* what they would be if he had met the conditions of his epistemic obligation—this is to say that p cannot be justified to a degree greater than 0.4 for A, and that q cannot be justified to a degree greater than 0 for A.

Examining the consequences of this treatment of A's evidence set, we may first note that A's epistemic relation with respect to p is ultimately unaffected by the existence of his unfulfilled epistemic obligation, since $r_A(E_A) \Rightarrow < p, 0 >$ in item 2 (*i.e.*, our obligation-independent assessment of A's evidence base) and thus the As-If

[10] To be clear, this only requires that an agent explicitly believe (*i.e.*, have in her active fragment) propositions which are contradictory, not that she explicitly believe the contradiction itself (although in virtue of the active fragment being closed under unrestricted classical implication we do say that she *implicitly* believes the contradiction).

[11] this is to say that if an agent believes x and also believes y, where $y = \neg x$, then for that agent $< x, 0 >$ and $< y, 0 >$, and neither x nor y can lend any justificatory support to the agent's other beliefs. If the agent believes propositions q, r, and s only on account of their support from either x or y, then the justificatory statuses of q, r, and s are all 0. Furthermore, if propositions t is believed in virtue of x and some undefeated proposition z, then t can only enjoy as much justification as it is provided by z.

Principle's does not exercise any influence over A's epistemic relation to p by capping its justificatory status at 0.4. A's epistemic relation to q, on the other hand, is severely impacted by the As-If Principle's treatment of epistemic obligations. Independent of our considerations of A's unfulfilled epistemic obligation, A's actual evidence base (item 2) yields the belief-justification pairing $< q, 0.7 >$, but because we treat A as though his evidence base is reflected by item 3, he is at most justified to degree 0 in believing q. If A is unaware of his epistemic obligations in this case, it seems likely that he would believe q since his actual evidence base fairly strongly supports q, but in virtue of A's unfulfilled epistemic obligation any such belief would turn out completely unjustified, according to the As-If Principle. We thus see that the application of the As-If Principle intuitively allows for decreases in agents' degrees of justification in cases where what an agent actually believes and what an agent is obligated to believe come into conflict. We may now move on to consider whether the As-If Principle allows for an agent to have a higher degree of justification for a belief than the baseline justification set by the conditions of that agent's epistemic obligation.

For our second example, let us assume that E_A is such that $e_1, e_2, e_3 \in E_A$, so that it should turn out that A's belief that p is justified to degree 0.9 and that A's belief that q is justified to degree 0. The concern that I intend to dispel is that, provided the baseline condition for A's satisfaction of C is that $e_1, e_2 \in E_A$, the As-If Principle implies that A's belief that p is justified *at most* to degree 0.7, on account of the assumption that when $e_1, e_2 \in E_A$ and $e_0, e_3 \notin E_A$, $r_A(E_A) \implies < p, 0.7 >$. Let us begin this analysis by substituting the relevant symbols from the example into the consequent of the As-If Principle, giving us "for any belief p that A holds in virtue of some method of reasoning over a set of evidence, $r_A(E)$, the justificatory status of p cannot exceed the justificatory status that p would enjoy if A had met O and had formed p in virtue of $r_A(E)$." We see that the basis for the present concern is that we may be bound by the As-If Principle to treat E as though it does not include e_3, but only those evidence items constituting C. If our obligations restrict our evidence base to only those items that we are obligated to include under the As-If Principle, then we certainly ought to abandon it, since it would unrealistically restrict our ability to generate well-justified beliefs.

Note, however, that in the definition of the As-If Principle the evidence set E is defined in relation to the beliefs that the agent *actually* forms, and, more specifically, is fixed as the relevant set of evidence that the agent has used in forming her actual beliefs. For the case at hand, then, we assume $e_1, e_2, e_3 \in E$ (and $e_0 \notin E$). The fear is that the As-If Principle does not allow the justificatory status of p to exceed that which p would have in virtue of $r_A(E)$, for E such that $e_1, e_2 \in E$ and $e_3 \notin E$, but this is not the case. The As-If Principle preserves the standing evidence base, and only performs its crucial function by adding in whatever obligatory beliefs have been neglected by the agent, using problems generated by producing contradictions within the agent's evidence base to nullify (or at least lessen) the justificatory status of certain of the agent's beliefs rather than directly restricting the evidence base. Because C does not directly limit the evidence base from which the agent's beliefs derive their justificatory statuses, this fear is misplaced.

Therefore, in virtue of e_1 and e_2 being elements of E_A, we see that the status of p produced by $r_A(E_A)$ is the status that p would enjoy if A had met O, because A *has* met O in virtue of holding e_1 and e_2 in addition to e_3. The existence of the obligation to meet O is irrelevant to the justificatory status of A's belief that p so long as it does not impose any novel beliefs upon A. In this case $r(E_A)$ produces a justification of 0.9 for p, and the As-If Principle is thus shown to allow for evidence beyond one's basic obligations to generate further justificatory weight for one's beliefs.

As the foregoing examples show, the proper application of the As-If Principle produces results that are aligned with our intuitions that unfulfilled epistemic obligations should be able to bring about normative defeat for certain of our beliefs (*i.e.*, lower their justificatory status), but that our epistemic obligations do not exert any destructive influence over our well-founded beliefs. This is precisely the intuition fueling Lackey's cases against the PVM, and thus accepting that these applications of the As-If Principle accurately illustrate the impact of epistemic obligations on the justificatory statuses of our beliefs strongly supports Lackey's counter-PVM position. As I mentioned in the course of the fist case, however, there is one crucial topic that requires further examination: how the contradictory beliefs posited by our applications of the As-If Principle bring about changes in the justificatory statuses of an agent's actual beliefs. This is the matter to which I will now attend.

3.7 The Justificatory Implications of Contradictory Beliefs

The As-If Principle requires that we treat agents who are epistemically obligated to believe particular propositions as though they do in fact believe those propositions. In the language of the SFM, we treat an agent with an epistemic obligation to believe p as though p is an element of her active belief fragment. Because agents' active belief fragments are assumed closed under unrestricted classical implication, an agent who is epistemically obligated to believe p and actually believes $\neg p$ is thus treated as though she believes $p \wedge \neg p$, and thus treated as though she implicitly believes all expressible propositions, *ex falso quodlibet*. In order for this to produce the effect desired by Lackey, we need to consider precisely how the As-If Principle's conflict-production treatment of epistemic obligations impacts the epistemic statuses of an agent's beliefs.

The solution that I propose invokes the notion of epistemic quarantining, whereby the existence of conflicting belief contents within an agent's active fragment prevents that agent from enjoying non-zero justification in any proposition which he is assumed to both believe and not believe. The idea is, essentially, that when an agent is taken to believe p and $\neg p$, she is unjustified in both of these beliefs,[12] given that

[12]One might object that we often find ourselves inclined towards contradictory beliefs, but that in such cases we are not wholly unjustified in both. This, however, conflates belief with certainty. David Lewis discusses the importance of this difference briefly in *On the Plurality of Worlds* (Lewis, David. *On the Plurality of Worlds*. Oxford: Blackwell, 1986.), when he considers "doublethinking" agents who are "simultaneously disposed toward [multiple] systems" of beliefs (31). Lewis points

each belief should rule out her confidence in affirming the other. So long as both beliefs remain within the agent's doxastic structure, both are denied the ability to contribute justificatory force to any of the agent's other beliefs, in virtue of the fact that the logical space that one of the beliefs tends to endorse and the other tends to reject is trapped in a state of deadlock. This is to say that in the case where I am taken to believe both p and $\neg p$, if I have a further belief that z which derives some of its justification from my belief that $\neg p$, then so long as I am assumed to believe p my belief that z will not receive any justificatory power from my belief that $\neg p$. This is the sense in which contradictory beliefs become quarantined: the propositional space, so to speak, over which the contradictory beliefs conflict becomes epistemically neutralized, such that nothing dependent upon the state of that space can be rationally inferred.[13]

Let us return to the voting case, for a more concrete example. Under the As-If Principle's approach to epistemic obligation, I am presumed to harbor both the epistemically obligatory belief that the Democrat needs 48 percent of the vote to win, as well as my actual belief that the Democrat needs 55 percent of the vote to win. This carves out the following belief structure: I believe that the Democrat will lose if she receives less than 48 percent of the vote, the Democrat will lose and not lose (alternately, win and not win) if she receives between 48 and 54 percent of the vote, and the Democrat will win if she receives 55 or more percent of the vote. The As-If Principle thus asserts that I am deadlocked when it comes to my tendency to affirm one way or another what happens when the Democrat has between 48 and 54 percent of the vote, thus making this region of conceptual space inaccessible for justified belief formation. What happens, then, when I learn that the Democrat received 51 percent of the vote? Because the deadlock brought on by my contradictory beliefs leaves me indisposed to produce a justified belief about the outcome of the election in this situation, the information that the Democrat has 51 percent can only be used to infer the outcome of the election once the deadlock has been resolved. The only justified election-relevant information that I have is that the Democrat will lose with less than 48 percent of the vote, win with 55 percent or more of the vote, and that she has 51 percent of the vote; none of these beliefs warrant any further beliefs about the results of the election. My only justified course of action is to remain agnostic about the outcome of the election, so the belief I form that the Democrat lost the election because she received 51 percent of the vote is unjustified. Not only does this

out that while these agents are not "whole-heartedly" certain of their conflicting inclinations, their beliefs must "plunge one way or the other" at any given time (31). In our applications of the As-If Principle, we treat agents as plunging both ways, so to speak, and thus we are dealing with a different kind of situation from those in which agents merely have conflicting doxastic inclinations.

[13]There is a sense in which assumptions that things are one way or another will produce justified inferences, but this is only insofar as one is examining the logical form of the assumption, and has no bearing on the actual world. If I am taken as believing "It is raining" and "It is not raining," I can trivially take either statement on as a conditional and determine its logical consequences within my belief structure, but this will amount to nothing more than an exercise in formal logic, and cannot stand as a method for producing beliefs *about the world*. In this case my belief that "If it is raining then it is wet," for example, does not allow me to justifiably conclude that it is wet, even though I can justifiably believe "'It is raining' and 'if it is raining then it is wet' imply 'it is wet.'"

method of examining contradictory belief structures fit the bill for Lackey's attack on the PVM, but seems intuitive as far as our notions of epistemic obligations and normative defeaters go.

When an agent holds contradictory beliefs, she is taken as incapable of establishing justification for any beliefs that depend on her doxastic conflict resolving one way or another. Contradictions within agents' belief structures thus work to restrict the epistemic status of those beliefs whose epistemic statuses would benefit from the contradiction resolving one way or the other. Assuming that this account of the justificatory implications of agents actively believing contradictory propositions is correct, then the As-If Principle supports Lackey's argument that normative defeaters are able to effect dramatic changes in the justificatory statuses of agents' beliefs. Unless there is a flaw in my account of epistemic obligations, we are thus committed to the failure of the PVM. Because the Memory Justification Principle entails the PVM, this is deeply troubling for the Memory Justification Principle. In the following section I will investigate whether the Memory Justification Principle can be salvaged, in light of the foregoing discussion.

3.8 Epistemic Obligations and the Memory Justification Principle

Translating the foregoing discussion of Lackey-style cases into the language of memory traces, we have seen that in cases of normative defeaters it is possible for an agent to sculpt a memory trace with respect to a mental state that represents an unjustified belief, but to have that same memory trace produce a highly justified belief upon activation, provided the disappearance of the normative defeater. This is clearly at odds with the Memory Justification Principle (hereafter MJP), since the MJP asserts that a memory-belief can be at most as justified as the belief with respect to which the relevant memory trace was sculpted. We may note, however, that normative defeaters are overriding, and not undercutting—that is, it is only while a normative defeater is operative that it imposes limitations on the degrees of justification enjoyed by the beliefs which it contradicts, and thus allows for the justificatory statuses of the affected beliefs to "spring up" upon the defeater's disappearance. When a belief has been defeated by an overriding defeater, we maintain that all of the evidence in that belief's favor continues to support the defeated belief, albeit to no effect. This is precisely the notion that allows for Lackey-style cases—the elimination of an overriding defeater allows for the justificatory status of the previously defeated belief to "spring up" to where it would have been without the defeater. We thus see that Lackey-style cases rely on the preservation of the *internal* justifications[14] contributed by a memory-belief's causal history in a way similar to what is described by the MJP,

[14]Internal justifications for a belief may be thought of as the justificatory weights conferred on beliefs in virtue of the believing agent's aptly reasoning over his evidence base. External justifications, on the other hand, are conferred on beliefs in virtue of things other than the agent's reasoning process, and thus accounts for things such as how well the agent gathered his evidence, whether the agent had any outstanding epistemic obligations, etc.

but emphasize the need for some additional qualification that takes into account that *external* influences on the justificatory statuses of beliefs (*e.g.*, normative defeaters) allow for the possibility that memory-beliefs might enjoy higher epistemic statuses than their original counterparts.

We may then say that the MJP only tracks the internal justification of contents through memory, but allow for the possibility that external influences on the justificatory statuses of beliefs may require further considerations. In light of these concessions, let us revise the Memory Justification Principle as follows:

> *Internalized Memory Justification Principle (IMJP):*
> Let s_1 be a mental state with respect to which a memory trace m is sculpted at time t_1. Let c_1 be the propositional content represented by s_1, and let j_1^i be the internal justificatory status of c_1 at t_1. Let s_2 be the mental state produced by the activation of m at some time $t_2 > t_1$. Let c_2 be the propositional content represented by s_2. Finally, let the operator $sig()$ be the memory trace signature included in c_2. We say that the internal justificatory status j_2^i of c_2 at t_2 is given by $j_2^i = p * j_1^i$, where p is the proportion of the propositional content c_2 that is also represented in $sig(c_1)$.

This revised version of the MJP only commits us to a preservation view of the *internal* justification of content through the memorial process, and is thus perfectly compatible with the fact that external justificatory forces can cause changes in the total justificatory statuses of beliefs in ways that contradict the implications of the original MJP. To commandeer Burge's slogan for preservative memory, we might say that memory preserves beliefs with their internal justificatory histories. While the IMJP produces intuitive results in many cases, there is a possibility that variation over time in the collection of memory-beliefs that an agent has access to might lead the IMJP to produce counterintuitive results. I will now turn to present two of these problem cases.

3.9 Two Problems for the IMJP

The first problem is closely related to the Sam and Sophie case raised by Christensen and Kornblith against Burge. Consider two agents, Sam and Sophie, who both currently believe the Pythagorean Theorem—they both understand exactly what it means, and they could apply it with equal success in any situation where knowledge of the Theorem would be relevant. We may even assume that Sam and Sophie are internally identical whenever they are recalling and applying the Theorem. Furthermore, neither agent has any recollection of how they came to believe the Theorem. There is, however, one significant difference between Sam and Sophie. When Sam sculpted his Pythagorean Theorem memory trace, it was at the end of a rigorous proof of the Theorem, during which he built up the proof from basic mathematical principles. When Sophie sculpted her Pythagorean Theorem memory trace, on the other hand, she was in a position that made her belief formation highly unjustified—

perhaps she barely heard someone whom she did not have any reason to believe to be knowledgeable about geometry say from across a crowded room, "The Pythagorean Theorem states that the length of the hypotenuse of a right triangle is equal to the square root of the sum of the two sides' squares." Whatever the particulars of their situations, we will assume that according to the IMJP Sam will enjoy a high degree of justification in his recollections of the Pythagorean Theorem, whereas Sophie will have a low degree of justification when she recalls the Pythagorean Theorem. What is counterintuitive in this case is that internally identical agents can have wildly different degrees of justification for identical memory-belief contents, purely on the basis of the causal histories of their beliefs. Let us call this the Internal Twins Problem.

The second problem is quite similar to the Internal Twins Problem, but its consequences are even stranger. Consider Steve, an agent who has two memory traces that were sculpted with respect to identical mental states, namely, mental states representing the Pythagorean Theorem. One of the memory traces was sculpted at a time when belief in the Theorem was highly justified, the other was sculpted at a time when belief in the Theorem was highly unjustified—we might imagine that Steve had a Sophie-like experience during high school, followed by a Sam-like experience during college. Provided that the memory traces themselves are indistinguishable from one another in all respects other than their causal histories, we may assume that at any time when Steve recalls the Theorem it is equally likely that he is activating the memory trace that, according to the IMJP, produces a highly justified belief in the Theorem as it is that he is activating the memory trace that produces a highly unjustified belief in the Theorem.[15] Any time that Steve uses the Pythagorean Theorem, then, it is equally likely that he is highly justified in his application of the Theorem as it is that he is highly unjustified in applying the Theorem, even though for all he can tell there is no difference from one use to the next. Whenever we have duplicate beliefs with causal histories that endow them with different justificatory statuses, we are essentially gambling on whether the memory traces corresponding to the justified beliefs will be the ones which activate. Let us call this the Justificatory Gamble Problem.

One important clarifying note for these problems is that for my purposes the relevant differences between the agents in both of these cases are their memory-beliefs' *degrees of justification*, not whether their memory-beliefs are *a priori* justified. The IMJP only describes to the transmission of degrees of justification through the memorial process, and thus insofar as my objective is to defend the IMJP in these problem cases I will only attend to the matter of whether the IMJP can be shown adequately intuitive in characterizing agents' degrees of justification in problem cases such as the Internal Twins and Justificatory Gamble Problems. I will also be assuming that none of the agents' recollections are accompanied by a sense of doubt about the content of their recollections, such that they may be treated as being *a priori* warranted in be-

[15] We might note that this assumption of equal likelihood is nonstandard according to the SFM's notion that more recently activated beliefs are more central (*i.e.*, more likely to activate under relevant queries) than older beliefs. Were we not specifically interested in creating a problem case, we would expect for the beliefs that Steve formed in college to be more central to relevant queries than the beliefs he formed in high school.

lieving that their recollections present them with previously entertained mental state content; that is, each is *a priori* entitled to $prev(PythagoreanTheorem)$ given their exposure to $recall(PythagoreanTheorem)$. With these constraints in mind, we may now return to our treatment of these problem cases.

We have at our disposal three options to resolve these problems. The first option is to try to further modify our theory of how justification is transmitted through memory by developing some new justification principle such that the Internal Twins and Justificatory Gamble Problems are no longer consequences of our theory of how memory endows certain mental state contents with certain degrees of justification. Our second option is to rely on the IMJP's permission of external justificatory forces to develop some explanatory story about how external justificatory forces prevent these counterintuitive justificational disparities from occurring.[16] The final option is to bite the bullet, so to speak, and accept that our understanding of human-like memorial faculties commits us to the Internal Twins and Justificatory Gamble Problems.

Considering these options, I am skeptical about the possibility of further revising the IMJP to deal with these sorts of cases. Given the fact that the contents of the states with respect to which memory traces were sculpted should not be thought of as logically interacting with one another while those contents are not active, I do not see a promising means for incorporating into our account further claims about the ways in which the elimination of "similarly relevant" memory traces impacts the justificatory standing of surviving memory traces. Perhaps the IMJP could be revised in such a way that it might handle these sorts of cases more successfully, but I am uncertain about this possibility and will not pursue it here. Neither do I believe that the second option can succeed in a plausible way. The reason why it is essential to account for memory in our theories about agents' doxastic structures is that we are simply cognitively incapable of engaging large portions of our beliefs simultaneously. The second option only allows us high degrees of justification in our beliefs if we can remember why we hold each of our beliefs, such that we would either have to be constantly reaffirming our entire belief structures or else almost constantly engage with unjustified (or at least justification-impaired) beliefs. This seems unattractive, and thus I will pursue neither of our first options. Unless there is some viable alternative that I have left unmentioned, the IMJP commits us to the Internal Twins and Justificatory Gamble Problems. I will now argue that the IMJP's commitments regarding these problem cases are not wholly counterintuitive, and that we should not resist the IMJP on these grounds.

The discomfort common to the Internal Twins Problem and the Justificatory Gamble Problem is that in cases where memory provides us with content but does not supply any details about the source of that content we find ourselves incapable of investigating whether that content is justified, such that situations in which our memory-beliefs are justified are phenomenally indistinguishable from situations in

[16] We might think of the following as one approach to this sort of storytelling: As agents forget how they formed their beliefs, they collect epistemic obligations to reform their foundational beliefs, such that agents who cannot recall why they believe certain propositions cannot enjoy high degrees of justification in those beliefs (in virtue of the As-If Principle).

which our memory-beliefs are unjustified. Furthermore, given that justification is causally transmitted, the justificatory statuses of our future beliefs may well be influenced by the justificatory statuses of our recollections, such no matter how epistemically virtuous our present cognitive activities are we run the risk of being unwittingly haunted by our past epistemic viciousness. This is just to say, however, that the IMJP provides thoroughgoing support for the general notion of virtue epistemology.[17] As epistemologists, we think of the utility of memory as deriving from its ability to allow for temporally extended belief development—we treat memory as though it provides us with "pause/play" buttons that govern our exposure to specific mental state contents, such that we might pick up an interrupted thought process almost exactly where we left off.[18] If we reembark upon a thought process that began viciously, then the causal history of our present acts of belief formation will be rooted in unjustified beliefs; likewise, reembarking upon a virtuously founded thought process will causally connect our recently formed beliefs to justified predecessor beliefs. The intuition behind virtue epistemology is, simply put, that good behaviors (epistemic or otherwise) lead to rewards and bad behaviors lead to consequences, and in this way the Internal Twins and Justificatory Gamble Problems serve to illustrate that there are no statutes of limitations for matters of epistemic justice. The moral of the Internal Twins and Justificatory Gamble Problems is not that the IMJP is flawed, but that insofar as we wish to be certain in our higher-order beliefs about our memory-beliefs' justificatory statuses, we must be attentive to whether we know how we came to form our memory-beliefs.

As a final note on this matter, we may recall that the SFM generally attributes greater centrality (dispositional probability to activate with respect to a query) to more recently formed and more frequently engaged beliefs. This means that if an agent initially had an unjustified belief in the Pythagorean Theorem, but recently performed a rigorous proof of the Theorem, then the more recently developed memory trace would be more likely to activate during a Pythagorean Theorem-relevant query. Assuming that our belief-forming abilities tend to become more virtuous over time, then considerations of centrality reveal that our more virtuously formed beliefs will tend to be privileged over their less virtuously formed counterparts. If the problem cases mentioned above seem more troublesome to some than they appear to me, we may note that reminding oneself of the justifications of one's most important beliefs from time to time will greatly lessen the probability of inadvertently activating any unjustified duplicates of those beliefs, and thus there is a practicable solution to the concerns that the Internal Twins and Justificatory Gamble Problems raise.

[17]Roughly put, the general idea behind virtue epistemology is that the degrees of justification enjoyed by agents' beliefs is derived, at least in part, from the degrees to which those agents behaved in intellectually virtuous ways in arriving at those beliefs.

[18]The crucial caveat to this claim would be role that the memory trace signature elimination inference plays in giving us access to recollected content. Provided my argument in the previous chapter, we are *a priori* entitled to content within the memory trace signature operator in the case that that content is presented without an accompanying sense of uncertainty, such that while our memorial processes do not strictly speaking return us to a past cognitive state they can very closely approximate such a return.

Acceptance of the IMJP does not open the floodgates to skeptical problems related to our uncertainty about the origins of our memory-beliefs—at least not any more than any other theory of memory will.

3.10 Concluding Remarks

Given that epistemic obligations can plausibly be said to effect the defeat of certain of an agent's beliefs, as I have argued, then it appears that the PVM is untenable in light of Lackey-style cases. The appearance and disappearance of epistemic obligations allow the justificatory statuses of an agent's beliefs to oscillate independently of that agent acquiring further evidence regarding those beliefs. Because these oscillations in epistemic status occur through time, the recurrence of the agent's beliefs will be a product of memory (at least for the most part), and this entails that memory-beliefs can enjoy varying degrees of justification depending on the presence or absence of relevant normative defeaters. This is to say that memory can reconstruct a belief that has a greater justificatory status at the time of remembering than it did at the time that the belief was established, independently of the agent acquiring any further evidence for that belief, and thus the PVM has been shown incorrect and must be abandoned. Furthermore, because the original Memory Justification Principle entailed the PVM, we discovered that it was necessary to restrict the claims made by the Memory Justification Principle to describing the transmission of *internal* sources of justification through memory. The Internalized Memory Justification Principle reflects this realization, and thus finds a stable compromise between the intuitions backing the PVM and the existence of epistemic obligations and normative defeaters.

Conclusion

At the outset of this thesis, I claimed that my overarching intention was to demonstrate that considerations of memory are indispensable in epistemology.

In the first chapter, we began by considering the argument that belief fragmentation is a useful (if not necessary) tool for developing adequate formal models of epistemic agency. Despite the theoretical advantages of developing a fragmentation-based model of agents' belief structures, the methods for implementing belief revision and update for these models ran into psychological plausibility concerns. I linked the source of these fragmentation models' problems to their use of ontologically robust manner of grouping beliefs into subject-specific fragments. In response, I sketched the framework an alternative model, the Single Fragment Model, which allows for agent's individual beliefs to join and leave a single active fragment at any time, depending on their relevance to the agent's cognitive context and their centrality with respect to other beliefs relevant to that context. Furthermore, because the Single Fragment Model uses changes in centrality scorings to account for belief revision, it does not encounter the same problems that other fragmentation models do; instead of having to rely on the sort of distinct merging and non-merging options for belief revision that threatened the psychological plausibility of the B-structures model, the Single Fragment Model allows for fluctuations in beliefs' centrality scorings with respect to relevant queries to perform the essential functions of belief revision without actually necessitating the removal of beliefs from agents' belief corpuses. We thus saw in the first chapter that the Single Fragment Model serves as a fragmentation-based model of our doxastic structures which derives its formal advantages primarily by approximating how memory behaves, which reveals that considerations of memory are crucial even for the most abstract models used to model cognition for formal epistemology.

The second chapter began by introducing Tyler Burge's theory of preservative memory, which Christensen and Kornblith fault for being at odds with the cognitive psychology commonplace that memory is a reconstructive process. While I determined that Christensen and Kornblith's argument fails to clearly show that preservative memory is incompatible with memory being a reconstructive process, this is primarily due to Christensen and Kornblith's underspecification of what they mean when they call memory reconstructive. In order to better understand how memory is reconstructive, I thoroughly examined the memory trace account of recollection (presented by Mohan Matthen), which uses non-contentful memory traces to explain how mental state contents appear to be preserved. Using memory trace theory, I developed the Memory Justification Principle as an intuitive means for understanding how

justification is transmitted through the memorial process. While the Memory Justification Principle seemed to be a useful first step for arriving at a verdict on the issue of preservative memory, we needed a further account of how the memory trace signature content could be inferentially eliminated; if this signature content could be *a priori* eliminated in certain kinds of cases, then we would know that in those sorts of cases preservative memory would be possible. Our understanding of this inferential pattern largely hinged on our understanding of the memory trace signature content appended onto recollected mental states, and considerations of the possibility (attributed to Alex Byrne) that the memory trace signature content must simply be analyzed as our recognition of the distinctly memorial mode of presentation. Eventually, we determined that memory trace signature content can be eliminated *a priori* at least in cases of forever-after propositions, and thus we saw that at least in cases where a memory trace behaves in a purely preservative manner and produces a mental state content whose content may be represented with a forever-after proposition Burge-style preservative memory may be said to have been employed. After showing that memory trace theory provides picture of reconstructive memory that is compatible with the notion of perfectly preservative memory, I argued that the apparent conflicts between memory trace theory and the Single Fragment Model are only superficial, and concluded by showing that the memory trace approach to defining latent belief actually provides strong support for the Single Fragment Model.

My final chapter begins by introducing the Preservation View of Memory as a consequence of my previously developed Memory Justification Principle, and as a belief that Lackey believes to have falsified. Lackey's argument requires that we grant the existence of normative defeaters, and appears to succeed if we grant this requirement, although no further support is given for the existence of normative defeaters. We then turned to investigate whether we are subject to certain epistemic obligations as epistemic investigators. After unpacking these notions with the help of Hilary Kornblith, we determined that there is nothing immediately objectionable about our having epistemic obligations to behave as responsible epistemic investigators, and, furthermore, with the development of the As-If Principle as a means for determining how unfulfilled epistemic obligations impact the justificatory statuses of our beliefs we were able to see that Lackey's cases against the Preservation View of Memory were successful. Luckily, not only were we able to change the Memory Justification Principle so as to only describe the transmission of internal sources of justification through memory (salvaging most of its utility), but we were able to notice that Lackey's cases against the Preservation View of Memory actually depended on an intuition along the lines of the Internalized Memory Justification Principle. The chapter concluded by discussing the seemingly counterintuitive Internal Twins and Justificatory Gamble Problems, which I argued are not actually as counterintuitive as they might appear, and are actually useful in showing that the Internalized Memory Justification Principle supports the general principle of virtue epistemology—that good epistemic behaviors warrant rewards and poor epistemic behaviors warrant consequences—in a thoroughgoing way.

In the course of these three chapters, I have demonstrated that it is crucial for our epistemic theories to account for memory. First, in virtue of abstractly approximat-

ing the processes involved in memory, the Single Fragment Model was shown to be a successful and psychologically plausible fragmentation-based model of our doxastic organization. Memory trace theory revealed the extent to which memory is a reconstructive processes, and through our discussion of the theory's nuances determined that preservative memory is compatible with memory being reconstructive and that memory traces allow us an intuitive means for determining when it is correct to ascribe latent beliefs to agents. Furthermore, our considerations of Lackey-style cases showed that it is possible for memory-beliefs to enjoy epistemic statuses superior to those of their original counterparts. Finally, we discovered that the Internalized Memory Justification Principle provides strong support for a thoroughgoing virtue epistemology, in virtue of how it treats the Internal Twins and Justificatory Gamble Problems. Most importantly, however, we can see that there is still much more to discuss on the matter of the philosophical significance of memory. It seems possible that when memory traces activate to produce perceptual experiences, we actually perceive objects existing in a different time slice of the world.[19] It seems possible that a theory of perception could be developed around sense-data triggering the relevance querying process prior to generating our sense perceptual experiences, such that conceptual content contributed by memory may be understood to saturate our perceptual experiences. Wherever things seem possible, we need philosophy to help us see what is impossible, and memory is a topic full of possibilities.

[19]Credit for this thought goes to Paul Hovda.

Works Cited

Burge, Tyler. "Interlocution, Perception and Memory." *Philosophical Studies* 86 (1997): 21-47.

Byrne, Alex. "Recollection, Perception, Imagination." *Philosophical Studies* 148 (2010): 15-26.

Chopra, Samir and Rohit Parikh. 2000. "Relevance Sensitive Belief Structures." 1-25. www.sci.brooklyn.cuny.edu/ schopra/maidone.ps

Christensen, David and Hilary Kornblith. "Testimony, Memory, and the Limits of the A Priori." *Philosophical Studies* 86 (1997): 1-20.

Egan, Andy. "Seeing and Believing: Perception, Belief Formation and the Divided Mind." *Philosophical Studies* 140 (2008): 47-63.

Kornblith, Hilary. "Justified Belief and Epistemically Responsible Action." *The Philosophical Review* 92 (1983): 33-48.

Lackey, Jennifer. "Memory as a Generative Epistemic Source." *Philosophy and Phenomenological Research* 70 (2005): 636-658.

Lewis, David. "Logic for Equivocators." *Nous* 16 (1982): 431-441.

——— . *On the Plurality of Worlds.* Oxford: Blackwell, 1986.

Matthen, Mohan. "Is Memory Preservation?" *Philosophical Studies* 148 (2010): 3-14.

McGrath, Matthew. "Memory and Epistemic Conservatism." *Synthese* 157 (2007):1-24.

Wittgenstein, Ludwig. *Philosophical Investigations.* trans. Ancombe, G.E.M.; P.M.S. Hacker; and Joachim Schulte. West Sussex: Blackwell, 2009.