VIGILANCE: SELF-CONTROL FOR PLANNERS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Samuel F. Murray

Ted A. Warfield, Co-Director

Robert N. Audi, Co-Director

Graduate Program in Philosophy

Notre Dame, Indiana

April 2019
People often pursue multiple goals and occupy overlapping social roles simultaneously. This occurs against the backdrop of strict psychological and physical limitations. In other words, we want to do a lot with a little. One way to navigate this tension is to scatter agency. This requires putting things on hold to pursue other opportunities at the right time, all while remembering to come back to previously undertaken endeavors when appropriate. This opens up the need for a capacity to manage psychological resources implicated in acting over time. In this project, I call this capacity vigilance and outline the role that it plays in managing first-order psychological capacities such as memory and attention.

I focus on the relationship between vigilance and other psychological constructs (attention, memory, self-control, and mind wandering) as well as normative concepts (responsibility). While the main aim of the current project is to illuminate the nature and function of vigilance, focusing on vigilance has implications for thinking about the structure of human agency. In particular, the theory of vigilance outlined here
explains self-control limitations without appealing to limited resource consumption. This also furnishes a more nuanced account of the role that motivation plays in action. Additionally, I conclude by outlining a robust empirical research program that falls out of the conceptual account of vigilance outlined here. Future work will follow up on this program in an effort to confirm the theory or identify potential areas of refinement.
CONTENTS

Figures............................................................................................................................... iv

Tables ................................................................................................................................. v

Acknowledgments............................................................................................................. vi

Chapter 1: The Nature and Norms of Vigilance
  1.1 Introduction................................................................................................................ 1
  1.2 The Importance of Planning...................................................................................... 2
  1.3 The Psychology of Planning Agents........................................................................ 3
  1.4 The Structure of Vigilance........................................................................................ 7
    1.4.1 Goals Structure Vigilance ................................................................................... 8
    1.4.2 Knowledge Structures Vigilance ........................................................................ 9
    1.4.3 Cares Structure Vigilance .................................................................................. 10
  1.5 The Function of Vigilance........................................................................................ 12
  1.6 Vigilance and Other Psychological Constructs....................................................... 17
  1.7 The Norms of Vigilance............................................................................................ 20
  1.8 Conclusion ................................................................................................................ 29

Chapter 2: Vigilance and Attention
  2.1 Introduction................................................................................................................ 30
  2.2 Vigilance and Attention............................................................................................ 33
  2.3 Attention Manifests Vigilance.................................................................................... 36
    2.3.1 Top-Down Attention............................................................................................ 37
    2.3.2 Controlled Attention........................................................................................... 39
    2.3.3 Automatic Attention............................................................................................ 42
    2.3.4 Bottom-Up Attention.......................................................................................... 45
  2.4 Performance Breakdowns......................................................................................... 47
    2.4.1 Examples of Performance Breakdowns.............................................................. 48
    2.4.2 Are Performance Breakdowns Memory Failures?.............................................. 52
    2.4.3 Are Performance Breakdowns Mechanical or Action Failures?....................... 55
  2.5 Attention and Performance Breakdowns.................................................................. 56
    2.5.1 The Cognitive Unison Theory of Attention....................................................... 56
    2.5.2 Structuring Theories of Attention...................................................................... 57
    2.5.3 Broadcast Theories of Attention........................................................................ 61
    2.5.4 Selection for Action Theories of Attention....................................................... 62
    2.5.5 Failures of Attention and Performance Breakdowns....................................... 64
FIGURES

Figure 6.1: Effects of a Two-Hour Watch in the Clock Task .................................................182
Figure 6.2: Effect of the Telephone Message in Clock Task ..............................................186
Figure 6.3: Effect of the Spacing of the Twelve Signals within Half-Hour Spells in Clock Task ..........................................................193
TABLES

Table 3.1 A Taxonomy of Vigilance Failures ........................................................................100
ACKNOWLEDGMENTS

First and foremost, I must acknowledge my debt to Eleonore Stump, Jon Jacobs, and Kevin Timpe, who cultivated my early ideas on responsibility and negligence that grew into this dissertation. Bill Newsome and Jon Cohen helped me wade through cognitive science literature on control and memory during early stages of this project. John Schwenkler deserves a shout-out for some very helpful suggestions he offered when I presented some of this material at the 2015 Tennessee Value and Agency conference.

Thanks to my dissertation committee members, Jeff Speaks and Sara Bernstein, as well as my advisers, Fritz Warfield and Robert Audi. I did not deserve the time they put into this work, and I hope that the final product lives up to their own Herculean standards of philosophical adequacy.

A number of people merit special mention. Chandra Sripada (also a committee member) gave critical feedback on different aspects of this material over the years. In many ways, he has served as a model for the kind of empirically-informed philosophy that I try to imitate here. Felipe De Brigard and Walter Sinnott-Armstrong hosted me at Duke for a semester when I was first putting these ideas down on paper, and Walter gave extensive, hand-written feedback on almost every chapter. They taught me how to do empirical moral psychology, and I would be utterly incapable of producing this
dissertation were it not for their patient guidance. Many thanks also to their students, Paul Henne, Vlad Chituc, Bryce Gessell, Aaron Ancell, and Greg Stewart for creating a collaborative, welcoming atmosphere at Duke.

I began to work seriously on vigilance in 2015 with Manuel Vargas. We spent many invaluable hours talking about responsibility and moral psychology, almost all of which fundamentally altered the direction of my thinking about a vast number of topics. Working with Manuel made me into a philosopher, and his influence spreads throughout this entire project.

Nearly final versions of every chapter were written in Bogotá, where I had the pleasure of working with Juan Pablo Bermúdez at the Universidad de los Andes. Juan Pablo was a valuable conversation partner, but he remains an even more valuable friend.

Two people deserve special mention. Kristina Krasich, my colleague at Notre Dame, taught me almost everything I know about cognitive psychology. Santiago Amaya taught me almost everything I know about crafting good philosophical papers. He played the role of mentor, critic, and friend at key moments. My life is richer for having known him, and not just because he is philosophically brilliant.

Finally, my father, Michael Murray, read drafts of this dissertation multiple times. He was my first philosophy teacher, and we talked almost every day about this project. Without his input, none of this would have ever happened.

And, to Laura Wells, thank you for your unconditional support and love as I worked on this dissertation.
CHAPTER 1:
THE NATURE AND NORMS OF VIGILANCE

1.1 Introduction

The use of plans enables the simultaneous pursuit of heterogeneous, complex goals. Human beings, for instance, frequently make plans as a way of managing the demands of labor and leisure. Some planners, such as human beings, must navigate this plural goal pursuit under strict cognitive limitations. This raises a problem for planning agents, namely how to coordinate and balance pursuing various goals in a coherent and efficient manner. I propose that vigilance, a disposition to become aware of goal-relevant information at the appropriate time, solves this problem of planning agency.

Sections 1.2 and 1.3 outline the importance of planning and the psychological components constitutive of planning agency. This sets up the Problem of Scarce Information. Vigilance provides one efficient and empirically plausible solution to this problem. Section 1.4 sketches the structure of vigilance, while Section 1.5 discusses the function of vigilance. Section 1.6 distinguishes vigilance from other psychological constructs, such as attention and memory. Finally, Section 1.7 investigates whether there are norms that govern exercises of vigilance. This chapter points to important and underappreciated issues about the psychology of planning agency, the nature of temporally extended action, and the norms that govern how one ought to balance
focusing on what one is currently doing with focusing on the goals one aims to realize in
the future.

1.2 The Importance of Planning

Do you want to go to Disney World? If you have a family of four and want to stay
inside the park during the summer months, then you need around $4000 for the trip
(Black, 2013). Want to get married? CNN reported that the average cost of a wedding in
2015 was $32,641. Of course, that depends on location. If you’re getting married in
Manhattan, you should plan to save around $82,000. Alaska, on the other hand, will run
you a cool 17 grand (Vasel, 2016).

Many of the big or exciting events in our lives require planning because they
involve the expenditure of money or time. But we also need planning to account for the
unexpected. The Centers for Medicare and Medicaid Services report that in 2015 the
average annual costs per person of healthcare in the US rose to $10,345 (Bloom, 2017).
That’s why many financial consultants suggest having at least three months’ salary saved
for medical emergencies (Grant, 2015). As Richard Bolles suggests in his best-selling
What Color is Your Parachute?, “In today’s world, you must always have a plan B up
your sleeves” (2018: 16).

These examples highlight some interesting features of planning. First, planning is
useful because there are a variety of goals and projects that people pursue. Being healthy,
secure, and traveling are all different kinds of goods that can be attained (and maintained)
in different ways at different times. But the temporally extended pursuit of various goods
does not by itself necessitate planning. What forces us to be planning agents is that we
have scarce resources available to deploy in pursuit of these various goods. Creatures with limited memory, attention, and control (among other cognitive resources) need to coordinate their activities in order to balance when, how, and to what extent they will pursue various goals in order to fulfill a suitable number of them. Planning facilitates coordinating these activities in a coherent and efficient manner.

Thus, planning is important for creatures that are saddled with strict cognitive limitations and pursue projects with complex satisfaction conditions. The self-help and financial management literature implicitly assume this function in discussions of planning, but a similar message comes out in philosophical discussions of planning. Michael Bratman (2007: 53), for instance, writes:

> There are substantial pressures for mechanisms that support coordination and organization...Such mechanisms, further, need to respect basic limits we can expect to characterize the psychology of such creatures. We can expect such creatures to have limited resources of time and attention for complex reasoning and to be limited in what they know about themselves and their world...A plausible strategy here would be to add capacities to settle in advance on complex but partial plans of action...[This creature], then, is a planning agent.

But while planning agency enables the pursuit of complex, variegated goods, planning agency also introduces novel environmental pressures that need to be solved. To see this, we should first consider what kind of psychological apparatus is needed to support planning agency.

1.3 The Psychology of Planning

When it is the case that some individual has a plan, as Bratman suggests, this entails that the individual has a future-directed intention (or a set of such intentions). Plans are directed toward the future, because the function of plans (as stated above) is to
settle or streamline deliberation at the moment of action. Partly, this is because of coordination pressures. It is difficult to coordinate with others if we never settle on what to do in the future. Also, many of our temporally extended actions are complex, and it is easier for us to settle certain aspects of a plan in advance of needing to execute that plan. For this reason, plans tend to be partial because we specify as much detail as is needed and fill in information over time as necessary. For example, one might plan to stop at the store on the way home without settling on a particular route.

When we form plans we do not always continue to think about them right up until the moment of action. Part of the usefulness of plans (for limited creatures like ourselves) is that we can form them and store them without consciously attending to them, thereby freeing up resources to pursue more locally salient activities. Thus, after forming and settling on a plan, we often store those plans and recall them only in the relevant action context.

This outline of the elements of planning agency implies that planning agents possess a number of psychological capacities. First, deliberation about future courses of action requires a capacity for imagination and simulation (including some form of working memory to store and compare various simulations). Second, there must be some capacity to form future-directed intentions. Third, there must be some storage device for these plans, at least when we do not need to consciously attend to the plan from the moment of plan-formation up to the moment of action. This allows us to form a number of plans without putting a burden on our limited cognitive resources. Fourth, there must be some kind of recall system for our plans comprised of a recognition device and some kind of cue associated with plans for purposes of future recall. Match detection between
the present action context and the associative cue triggers plan recall. The fifth element, then, is an implementation capacity that implements the recalled plan and initiates action.

This rough sketch describes what psychologists call ‘prospective memory’ (see McDaniel & Einstein, 2007), and the term refers to the network of processes and states implicated in storing and recalling plans (the psychology literature uses the term ‘delayed intentions’ to refer to plans; see, e.g., Scullin, McDaniel, and Shelton, 2013). In a nutshell, then, planning agency requires the development of prospective memory and the associated mechanisms of prospective memory.

The need for prospective memory as part of planning agency introduces a unique environmental pressure, what I call The Problem of Scarce Information. Achieving a goal requires a certain level of attentiveness to those considerations that are instrumentally relevant to achieving the goal. But, as noted earlier, we form many goals over the course of our lives and we do not form new goals only when we have finished previously adopted goals (this, recall, was the reason for utilizing planning to connect pursuit of various goals together efficiently and coherently). These goals, collectively, would require more attentiveness than we have available to deploy. Hence, we need to scatter our agency in pursuing different goals in different ways over time (or, in other words, we scatter agency to engage in differential plural goal pursuit). Scattering agency, however, raises a new problem. We do not typically have in mind all the information necessary to follow through on all of our plans. The phrase ‘in mind’ refers to the state of having information within the scope of one’s occurrent awareness and accessible for

1 The idea of scattering agency is inspired by Roy Sorensen’s notion of scattered events, or events that have at least some non-overlapping temporal parts (see Sorensen, 1985).
consumption in a wide range of cognitive processes. This is the Problem of Scarce Information, and it describes the condition of a creature that utilizes planning and has scarce cognitive resources to deploy in service of planning.

The solution to these new pressures requires a certain kind of self-regulatory system that manages the various informational demands of one’s current activities with the informational demands of one’s future planned activities. To solve the problem the system must fulfill two requirements. First, the system needs to regulate what information is and is not in mind. Ideally, if the mechanism is functioning perfectly, then it will be the case that all and only relevant information is in mind. This entails both that new information is brought into mind when it becomes relevant and that information is filtered out of mind only when it is no longer relevant. The second is that this regulatory system is not limited in the same kind of way as the regulated capacities or processes.

Consider that limited creatures like us need some system that manages interactions between various cognitive processes in a way that respects the capacity limits of those processes. If, however, the limitations of the regulatory system derive from something that also grounds the capacity limits of the regulated capacities and processes, then an additional management system is needed that governs the interactions between these limited capacities.² At some level, you need a management system not subject to

² Why think that there need to be management systems at all? That is, why not have everything set up to work harmoniously without any executive functions? The answer is that management systems solve a dilemma. On the one hand, it is computationally cheap to establish a number of systems that perform computations that are well defined. However, these heuristic systems cannot dynamically adjust to novel environmental situations. Also, from an evolutionary standpoint, while this heuristic set-up is computationally cheap in the short-term, the inflexibility means that adaptation requires adding new systems, which is metabolically costly (see Anderson, 2014: 30). On the other hand, you could have a flexible system tuned to a broad, heterogeneous class of environmental stimuli, but this comes at the cost of processing speed. A system that incorporates managerial mechanisms that compute executive functions has
limitations similar to those of the managed capacities and processes. Thus, the second requirement is needed to block the threat of an infinite regress of meta-cognitive self-regulation systems. In Section 1.5, I return to the issue of whether there is any system that fulfills these dual requirements.

I use the term ‘vigilance’ to refer to the self-regulatory system that solves the Problem of Scarce Information. The label is partly stipulative, but the decision stems from the fact that vigilance, in ordinary usage, is a state of keeping watch. If we think of vigilance as a kind of watchfulness related to planning, then the label makes some intuitive sense. In this context, then, vigilance is a higher-order disposition to manage the distribution of psychological resources over time.

In what follows, I outline the nature of vigilance and the norms that govern vigilance, starting with some structural characteristics of vigilance. This will clarify the functional profile of vigilance, help to distinguish vigilance from other psychological constructs, and provide a basis for the norms of vigilance.

1.4 The Structure of Vigilance

While a substantive discussion of vigilance requires seeing the way in which vigilance interacts with different elements in an individual’s psychology, there are certain facts about vigilance that we can settle in advance. It might seem odd that we don’t start out with some account of the function or purpose of vigilance, though I think we can...
work with an intuitive notion of vigilance as some kind of power of noticing or focusing. Discussing the role that goals, knowledge, and cares play in structuring vigilance provides a basis for sketching the functional profile of vigilance.

1.4.3 Goals Structure Vigilance

The first structural claim is that goals structure vigilance. Consider the differences in vigilance between a basketball coach and a fan at a basketball game. The two participate in at least one shared activity, namely watching the game. But the vigilance profile of the two differs substantially. A good coach will reliably watch for and notice considerations related to, among other things, strategy and substitution and be expected to maintain awareness of these considerations throughout the course of the game. A fan, on the other hand, may notice only a subset of these considerations throughout the course of the game, and there is no expectation of maintained awareness. The relevant difference seems to result from the different goals of the individuals. The coach aims to coach, which thereby demands vigilance toward a distinct class of considerations, while the fan aims to be entertained.

Some of this difference might seem to result from different kinds of knowledge or experience. But perhaps not. Consider, for example, the difference in vigilance between an educated spectator (or television commentator) and a coach (e.g., Doris Burke and Tara VanDerveer). The commentator would reliably watch for and notice considerations relevant to announcing the game, while the coach reliably watches for and notices considerations relevant to coaching and winning the game. Those considerations might
overlap to some degree during the course of the shared activity, but they will also likely
diverge in certain places. For instance, during timeouts, the commentator might focus on
certain statistics or personal stories about the players that enhance the viewing experience
for the television audience, while the coach will focus on devising an effective strategy
for the next part of the game. Here, the difference in vigilance does not seem to result
from any difference in knowledge (assuming that Burke and VanDerveer have similar
expertise with respect to basketball), but a difference in goals.

1.4.3 Knowledge Structures Vigilance

Knowledge also structures vigilance. Consider the difference between a superior
and inferior coach (or, if you like, a superior and inferior chess player). The knowledge
that the superior coach or player has of the relevant activity enables perception of various
patterns or likely future events that might otherwise go undetected by others that lack
such knowledge. This might be due to stronger associations between a perceived event
and the likely consequences of that event. Whatever the precise details of the structuring
end up being, differences in vigilance between these two individuals (whether coaches or
players) likely result from differences in knowledge.

There is another dimension along which knowledge makes a difference to an
agent’s manifestations of vigilance. Take two individuals, each of whom is in a
relationship with a different partner. We can assume that each individual has the
happiness of their own partner as a goal. However, the means to achieving that goal
might be different depending on individual differences among the partners. And
knowledge of those differences will lead to different distributions of vigilance among
individuals. As a simple example, consider that one partner might value birthday celebrations more than anniversary celebrations, whereas the other partner values anniversary celebrations more than birthdays. Those differences would lead (or should lead) the respective individuals to attend to different sorts of considerations (birthday-relevant or anniversary-relevant, as the case may be) despite having the same goal at a certain abstract level of description.

1.4.3 Cares Structure Vigilance

The final structural elements of vigilance are an agent’s cares. Roughly, when an individual cares about something or someone, she has a desire to be in the presence of that person or thing and has some policy that aims to promote that thing or the good of that person. Additionally, when an individual cares about something or someone, her desire for that person or thing tends to trump other conflicting desires.\(^3\) This brief description of care aligns with recent accounts of the role of cares in practical agency (see especially Sripada, 2016-a).

Importantly, there seem to be dispositional effects of caring on vigilance. For instance, caring about one’s family means that you are disposed to allocate attention to them and their needs at the expense of allocating attention to other people or projects. If you care about being funny, you’ll focus on finding opportunities to insert witty comments. Conversely, people take lack of focus or attention as evidence of a lack of

\(^3\) This is meant to indicate that cares partially constitute the fundamental level of an individual’s hierarchy of desires. There are complex questions about how desires conflict with each other and how strong the trumping tendency must be in order for some desire to count as a care. However, these questions are orthogonal to the main issues of this paper, so I will set them aside for now.
care (think: “You never pay attention to me”). This kind of inference displays the inherent connection between caring and vigilance.

There might be a temptation to think of cares as goals, such that the two have similar effects on one’s vigilance. There are connections between goals and cares, but the main difference seems to be that goals have well-defined terminal states that determine standards of achievement, whereas cares lack these terminal states. Additionally, we can infer a substantive difference between goals and cares from the differences that each makes on an agent’s vigilance. Consider two individuals that have, as a shared goal, throwing a party. Assume further that both have the same knowledge and experience with respect to throwing parties. However, one individual notices when others are alone and in need of social interaction, whereas the other individual focuses on food and refreshment management (clearing out empty plates or getting refills). This might reduce to a difference in cares (one person cares more about the social interactions of the guests, whereas the other one cares about party atmosphere). Thus, when we hold fixed goals and knowledge, we can still generate differences in vigilance that result from agential cares.

Though we’ve focused so far on the way in which goals, knowledge, and cares come apart, it will often be the case that these elements interact and inform each other in structuring vigilance. Cares will inform the kinds of goals that one adopts, and pursuing one’s goals can affect the contours of one’s cares. Additionally, knowledge (or the acquisition thereof) can function as a kind of feedback loop that facilitates updating or revising cares and goals due to increased information. This forms the core network of certain aspirational or normative elements of one’s self-conception (see Higgins, 1987; Murray & Vargas, Forthcoming: §6).
1.2 The Function of Vigilance

The structural considerations in Section 1.4 lead into thinking about the function of vigilance. The structure that goals and cares provide to vigilance indicates that vigilance is a capacity for awareness of considerations relevant to one’s goals and cares. Knowledge affects the kinds of considerations that are accessible and the relationship between one’s cares provides some way of determining how to allocate awareness toward considerations related to different goals.

Exercises of vigilance (in appropriately configured environments) trigger the acquisition of relevant representational and motivational states. The standard of relevance for some consideration is set by one’s goals and cares. That is, the degree to which some consideration is instrumentally relevant to realizing a goal or acting on some care is directly proportional to the degree to which an agent’s vigilance is tuned to that sort of consideration.

We should not lose sight of the connection between vigilance and planning. Given that goals are terminal states of plans (and, thereby, components of plans), the relationship between goals and vigilance is an aspect of the broader relationship between plans and vigilance. Recall also that the purpose (or importance) of planning rests on our need to pursue heterogeneous, complex goods for the sake of having meaningful and fulfilling lives. Our cares relate us to those things that we take to be importantly connected to (or constitutive of) those goods. So, insofar as vigilance is responsive to our cares, vigilance is also necessary for relating us to and facilitating pursuit of the goods that we take to be necessary for living a meaningful life. Hence, both the goal-directed and care-directed elements of vigilance relate to the way in which vigilance facilitates
planning. Further, this supports the claim that vigilance is not merely accidentally related to planning, but essentially connected to planning and planning agency.

We can discern other aspects of vigilance’s function from the above discussion. First, vigilance facilitates monitoring the environment in ways that support the exercise of temporally extended agency. Given the formation of some plan, agents need vigilance to monitor the environment for cues that signal when to act on that plan or when they need to fill out previously unspecified components of the plan. Also, monitoring figures into an agent’s ability to update or change her plans given new information. When engaged in complex, temporally extended actions, agents sometimes revise or revoke their plans in light of new information. Without vigilance, agents would have difficulty monitoring their environments in cognitively efficient ways for new information that would suggest revising, revoking, or persisting in their plans.

Second, vigilance enables agents to implement their plans at the appropriate time and place. Successful exercises of vigilance bring about action-relevant psychological states that figure in the causal production of intentional action. Hence, the activity of vigilance issues in the successful implementation of an agent’s plan.

Finally, vigilance affords a kind of maintenance or regulation of complex, temporally extended action. Given that many of our actions are temporally extended, agents must be able to sustain and persist in the actions that contribute to realizing one’s plan. Vigilance, then, enables agents to follow through on implemented intentions, which, in turn, allows for successful goal-directed temporally extended action.\(^4\)

---

\(^4\) This might seem, to some, like armchair faculty psychology (a pejorative label, in my estimation). However, there is good empirical evidence that identifies a plausible neurobiological realizer
Vigilance seems to be limited for creatures like us. The question, posed in Section 1.3, is whether the limitations on vigilance derive from the same source as the limited capacities that vigilance regulates. This issue takes us beyond the scope of the present chapter, but I sketch one response to this question here. Roughly, capacity limits on psychological capacities such as attention, memory, and control derive from representational overlap, or multiplexing (see Allport, Antonis, and Reynolds, 1972), whereas vigilance limitations derive from motivation and opportunity cost calculations.

Consider the fact that we utilize task representations to guide goal-directed behavior. So, we store task representations that correspond to crossing the street, whisking the eggs, and watering the garden. However, you wouldn’t want a unique representation for every possible fine-grained task you can perform. For example, you don’t need a separate street-crossing task representation for every street you happen to cross. A single, generalized ‘street-crossing’ task representation will do (see Rougier et al., 2005).

This is true in general. Having a basic stock of representations that are suitable for multiple tasks is more efficient than having a highly specialized basic stock of representations.\(^5\) Utilizing a basic stock of generalized representations that can be flexibly

---

\(^5\) This claim is restricted to creatures, like us, with highly circumscribed representational capacities.
deployed across a variety of tasks is known as multiplexing. Multiplexing, however, introduces the possibility of channel cross-talk. If you have a variety of available input-output task mappings subserved by the same representation, then these tasks might potentially interfere with each other (Forbus, Gettner, & Law, 1995).

The Stroop task provides a simple example of this. In the Stroop task, a participant is presented with a word stimulus internally colored by a color stimulus (e.g., the word ‘red’ filled in with the color GREEN). There are two conditions: congruent and incongruent. In the congruent condition, the semantic value of the word stimulus matches the color stimulus (e.g., the word ‘red’ is filled in with the color RED). In the incongruent condition, the semantic value of the word differs from the color stimulus. The participant is told to name the color. Reaction times in the incongruent condition are higher relative to the congruent condition. One explanation for this is that the input-output mappings that correspond to word naming and color identification both utilize the same representation. Hence, the processes that subserve these two tasks cannot be activated simultaneously and activating one implies performance deficits for the other. This reveals that multiplexing correlates negatively with the capacity to multitask. Increased multiplexing increases the possibility of channel cross-talk. This reduces the capacity to multitask without interference (see Feng et al., 2014 for a computational model of the absolute limits of multiplexing on multitasking in an optimal control network independent of network size and number of control nodes).

Dramatic limits on our psychological capacities reflect the brain’s preference for efficient coding through multiplexing. We can infer the brain’s preference for multiplexing over multitasking from the fact that we can perform so few controlled tasks
simultaneously. This, however, generates a problem of interference. A high degree of
pathway overlap implies a greater number of potential sites of interference between task
mappings. To solve this problem of interference, you need a central manager.

The function of the manager unit is to adjudicate conflicts between overlapping
pathways and bias lower-level information processing in ways that support goal-directed
behavior (see also Botvinick et al., 2001). Biasing is accomplished by reconfiguring
information processing pathways to alter the threshold for a neuronal population to fire.
Hence, conflict management requires reconfiguring various networks that mediate
different task sets. The principles that govern such reconfiguration are cost-benefit
calculations sensitive to the latent reward structure of available tasks and the costs of
pursuing that task relative to further exploration (Momennejad et al., In Revision). So, the
source of capacity limits for attention, memory, and control stem from the fact that the
processes these capacities subserve draw from the same representational pool, whereas
limits on vigilance derive from shifting motivation due to sensitivity to changing
opportunity costs afforded by the organism’s environment. Again, this issue requires
separate, focused treatment, but the discussion here illustrates some evidence that
vigilance is not limited in the same way as the capacities that vigilance regulates.

The exercise of vigilance manifests in the formation, acquisition, or
implementation of action-relevant representational or motivational states. Thus, when one
exercises vigilance, this brings about appropriate beliefs and desires (or triggers the
manifestation of particular dispositions to form certain desires and beliefs) that lead to the
formation or acquisition of relevant proximal intentions. In other words, exercises of vigilance activate motivating reasons that figure in the causal production of intentional action. Vigilance mediates between the world and the mind of the agent.

1.3 Vigilance and Other Psychological Constructs

At this point, it is useful to consider the relevant differences between vigilance and other psychological constructs, such as attention, memory, and control. At certain points, vigilance seems like attention (note that the structural and functional elements of vigilance involve noticing, awareness, and attending, which all sound like attentional activities). Additionally, it seems that vigilance does not add anything to our cognitive economy over and above the conjunction of attention, prospective memory, and cognitive control. What does vigilance itself add to our psychology?

To distinguish these constructs from each other, consider first three cases of planning errors (these are taken from Amaya & Murray (Submitted)).

**Skipping.** You want to check out a book from the library and do some grading in the evening. The library is on the way to your car, so you decide to get the book on your way home. You write the call number down and put it in your bag. Later in the evening, when you pull the books out, the call number falls to the floor. Only then do you realize that you drove straight home without stopping at the library.

**Derailing.** You want to check out a book from the library and do some grading in the evening. When you get to the library, you run into a colleague. She asks you about the memo sent a few days ago announcing changes in the faculty handbook. Given what she says, you realize you need to reread the memo in preparation for the upcoming faculty meeting. You drive home and spend the evening reading and thinking about the

---

6 I do not mean that beliefs and desires exhaust the range of cognitive and conative states connected to vigilance. ‘Belief’ and ‘desire’, here, stand in for the whole set of action-relevant motivational and representational states that we might acquire.
memo. In the morning, as you grab your bag on the way to work, you see the exams and realize you never did any grading.

Persevering. You want to check out a book from the library and do some grading in the evening. When you get to the library, you go straight to the shelf to get the book you need (you want to avoid running into anyone so that you don’t get home too late to do some grading). Once there, however, you decide to skim the book to see if it is worth taking it home. This takes more time than anticipated, and by the time you get home, you realize it’s too late to do any grading.

One might be tempted to see these planning errors as merely memory failures, lapses of attention, or control failures. But succumbing to this temptation is a mistake. Look again at skipping. In that case, you go straight home from the office without getting the book you need. This might seem like a matter of attention, but that characterization is misleading. A realistic version of Skipping probably includes the person thinking about something else. Perhaps it’s running through a mental grocery list, or thinking about an upcoming presentation, or wondering whether it’s time to encourage your kid to start taking piano lessons. The content of the thoughts are irrelevant. The point is that there’s something that acts as the object of your attention. Those things might be perfectly relevant to your overall life plan or commitments. So your attention is directed at relevant things. It’s just not appropriately allocated relative to a plan adopted earlier that is designed to achieve a variety of different goals. Thus, when skipping, you attend to something, just not the right thing relative to your plans. The attentional failure is not fundamental; rather, the failure is derivative on a misallocation of attention.

In derailing, you make it to the library, but your colleague prompts you to look again at the faculty handbook in advance of an upcoming meeting. It seems that your memory fails you here, but that’s not the case. You do remember some things, like the
handbook and the meeting. You just don’t remember the right things relative to your previously formed task set.

When persevering, you get to the library, but spend too much time reading the book. This might look like a failure of control, but it’s not. Finding the book in the stacks, reading through the book, and assessing the importance of it to your current research all require some measure of control. You simply fail to allocate control appropriately or to the right activities.

There is a common pattern here. In each case, some psychological capacity is engaged and functioning normally relative to the task in which it is engaged. But the capacity is not deployed in the right way. This gives the appearance of these failures arising from these capacities failing to perform their functions. But the current discussion shows that these failures arise from the mismanagement or misallocation of these capacities, not from improperly functioning lower-level capacities. As vigilance manages lower-level capacities, the failures are fundamentally failures of vigilance.

Importantly, however, this discussion shows that the activity of vigilance cannot be reduced to the functional profiles of attention, memory, and control. All of these capacities must be appropriately managed relative to the structure of one’s plans. This management does not fall under the purview of these capacities or any set of them. A higher-order capacity is needed. There is a unique role for vigilance within the career of a planning agent.
1.4 The Norms of Vigilance

The structural and functional considerations thus far provide some basis for thinking about the norms of vigilance. We think agents can be criticized in light of deficient exercises of vigilance. In general, we criticize agents for forgetfulness, inattentiveness, and other forms of negligence even when these failures do not reflect the underlying character of the agent (see Murray et al., 2019; Smith, 2005). This is some evidence that there are such normative standards.

Some agents also seem to systematically display sub-optimal vigilance, as in cases of ADHD, forms of autism, or scrupulosity (cf. Summers and Sinnott-Armstrong 2015: 952, 959). So, to provide a full account of vigilance, one must provide an account of the norms that govern vigilance. That is the task of this section.

When we consider the various ways in which vigilance breaks down or the ways in which agents display sub-optimal levels of vigilance, we can get a clearer pictures of the norms of vigilance. First, consider the sub-optimal vigilance of ADHD and autistic individuals. Recent work on the neurobiological factors of these clinical conditions reveals that ADHD and autistic individuals each display a distinctive vigilance deficit. Autistic individuals are highly attentive to task-relevant considerations, but display almost no sensitivity to salient, unanticipated considerations that are relevant to plan updating. ADHD individuals, on the other hand, are highly sensitive to novel stimuli, but display diminished sensitivity to task-relevant considerations (see Fogelson et al., 2014; Gonzalez-Gadea et al., 2015; Gonzalez-Gadea et al., 2016; Lawson, Rees, and Friston, 2014). These studies show two different poles of vigilance deficits. Autistic individuals manifest monitoring deficits insofar as salient stimuli fail to be detected and issue in plan
ADHD individuals, on the other hand, manifest maintenance deficits in failing to follow through on plans insofar as being vigilant toward plan-relevant considerations is constitutive of following through on one’s plans.

These two examples reveal an instrumental rationality associated with vigilance. That is, insofar as vigilance facilitates planning, vigilance ought to contribute to the realization of one’s plans. Vigilance that falls short of making this contribution is proportionally substandard. But notice that this normative constraint on vigilance is just inherited from the kind of instrumental rationality that governs acting on one’s plan (see Bratman, 1987: 107-10; Audi, 1991: 371-72). Therefore, this normative standard is not unique to vigilance. And, in fact, many of the normative standards that govern vigilance are derivative in this way.

Before getting to specific standards, we should draw a helpful distinction between basic and secondary norms. Basic norms are such that one has a normative reason to comply with the norm simply in virtue of understanding the norm. For example, prohibitions against murder and adultery are primary norms, meaning that one has a normative reason to refrain from both murder and adultery simply in virtue of one’s understanding the norm that prohibits these activities. Derivative norms are such that one has a normative reason to comply with the norm only in virtue of (at least some subset of) one’s prior commitments. So, I might have a normative reason to comply with a norm

---

7 This does not mean that the rationality that we attach to acting on one’s plans is merely instrumental. Bratman’s picture is complicated in two ways. First, he thinks that we can assess plans from either an internal or an external perspective. The internal perspective assesses the rationality of acting on one’s plans instrumentally. The external perspective incorporates considerations of instrumental rationality, but includes other considerations like norms of nonreconsideration and an assessment of the goals and desires that inform the formation of one’s plans.
that prohibits double dribbling with a basketball, but only in virtue of some prior commitment I made to play the game of basketball.\textsuperscript{8}

Vigilance, I suggest, is largely governed by these derivative norms. This makes sense, in part because the capacity for vigilance is structured by underlying dispositions of an agent. These dispositions correlate to commitments, cares, knowledge, and goals of an agent that then provide normative constraints on exercises of vigilance. And agents have a normative reason to exercise vigilance in specific ways only given their antecedent commitments to inhabit various roles (whether socially or intrapersonally defined) or other kinds of interpersonal relationships.

Additionally, because vigilance supports and regulates planning, the same sorts of normative standards that govern intentions (or the formation of these psychological states) will also govern exercises of vigilance. For example, if there are rationality and consistency constraints on intentions, then these constraints will also bear on exercises of vigilance. So, we might posit the following weak consistency constraint on exercises of vigilance:

\textit{Weak Vigilance Consistency}: If distal intention $I$ and distal intention $J$ require awareness of considerations $c$ and $d$, respectively, then Agent should be vigilant with respect to $c$ and $d$ only if awareness of $c$ does not preclude awareness of $d$ and awareness of $d$ does not preclude awareness of $c$.

For example, watching the football game on Sunday and watching the kids might be inconsistent, where inconsistency between intentions can be understood either as global

\textsuperscript{8} One might think this is a trivial point because all norms are derivative in virtue of supervening on contractual arrangements (e.g., Rawls, 1971 and Scanlon, 1998). However, even contractualists can accept the division between basic and derivative norms, as contractualist views appeal to basic norms like a principle of fairness or a principle of rational co-deliberation.
inconsistency between intentions and an agent’s beliefs and desires or as intentions whose corresponding actions cannot be jointly performed). If this is the case, then one should not intend to do both activities and one should not be vigilant toward both activities.

Here, however, we derive the consistency constraint on vigilance from the more fundamental consistency constraint on intentions (cf. Bratman 1987: 31-32). That is, agents should not form inconsistent intentions (whether this is understood as intentions that are inconsistent with an agent’s beliefs or intentions that cannot be jointly realized), and the consistency requirement for vigilance reflects this normative standard. There is, however, a role for the consistency requirement on vigilance, as some inconsistencies between intentions might reveal themselves only at the level of demands on vigilance. The inconsistency between the intention to watch football and watch one’s kids is an example of this.

Therefore, whatever normative constraints bear on intentions will also bear on vigilance (the same goes for belief, desire, and other action-related psychological states). The basic principle that supports this claim is that exercises of vigilance should not contribute to inconsistent or irrational plans. The tight connection between vigilance and planning (and, by extension, vigilance and intention) supports this claim.

I do, however, think that there is a normative standard unique to vigilance. To see this, consider the case of individuals with the clinical condition of scrupulosity. Scrupulous individuals are those with obsessive-compulsive disorders about morality or practices that take on a moralized character in light of an individual’s ideological commitments. So, instead of having an obsession with checking locks or washing hands,
these individuals have an obsession with various moral or quasi-moral practices, such as prayer. For example, consider the following case description of a scrupulous individual:

The bar mitzvah at 13 years of age represents the passing into manhood, when the Jewish males becomes responsible for his own religious practice. At this time, [Ezekiel] started spending excessive time in prayer and its preparations. His prayers took up to three hours daily, about three times longer than his peers. Despite the time spent, the content was abbreviated, less important parts being omitted, so that the significant sections could be said at the correct time. When he said the most important sentence, he doubted if he had complete faith at that moment, or he was assailed by a desirous thought of a female neighbour. He would go back over this line in his mind for up to 10 minutes. Similar difficulties over another section would cause this prayer to take up to 20 minutes, at least four times the average...He considered his rituals to be excessive, but experienced no resistance to carrying them out, and said that he would have put up with them were it not for his wife, who was contemplating divorce (Greenberg, Witztum, & Pisante, 1987: 33).

Notice, however, that the scrupulous individual doesn’t manifest vigilance deficits that are equivalent to ADHD or autistic individuals. There is neither a monitoring failure nor a failure of follow-through. In many ways, the scrupulous individual manifests exemplary monitoring and resoluteness. Some, however, might think that scrupulous individuals display suboptimal planning, or a deficit in some cognitive process other than vigilance. So why think these individuals manifest distinctive vigilance deficits?

To answer this, we need to discuss further what scrupulosity is. Scrupulosity is a form of obsessive-compulsive disorder (OCD). OCD is an anxiety disorder characterized by persistent, intrusive thoughts that the individual reduces through repeatedly performing a behavior. This behavior is considered the compulsion. While many people experience intrusive thoughts, most just ignore or reframe these thoughts (Salkovskis and Harrison, 1984). OCD individuals, however, fixate on and worry about these thoughts.
rather than ignoring them (Wells & Davies, 1994). Scrupulosity is a form of OCD involving moral or religious compulsions.

As Summers & Sinnott-Armstrong (2015: 951) explain, there is a crucial difference between scrupulosity and obsession with religion:

Someone who is obsessed with a religion…will care deeply about the features of that religion: practicing its rituals, studying its dogma, learning about the lives of its exemplars. She may be a religious leader, teacher, maybe even a saint. A scrupulous patient, on the other hand, will have religious anxieties and religiously-flavored compulsions, but may not care about the features of that religion per se, or may not give appropriate weight to the religion’s various features…The scrupulous Catholic, [for example], prays the rosary as a way of keeping bad thoughts or events from happening.

The comment about ‘appropriate weight’, in particular, is crucial for seeing scrupulosity as a manifestation of suboptimal vigilance, and not attributable to some other cognitive deficit. The scrupulous individual improperly balances various activities. This shows that the problem occurs at the level of regulating various cognitive processes, not at the level of the processes themselves.

Further, scrupulosity-based compulsions seem to occur contrary to one’s plans. Consider the story of Ezekiel from before. He “considered his rituals to be excessive” and that he “would have put up with them”. A natural interpretation of the case is that Ezekiel did not plan or commit to lengthy prayers, which explains why he comes to view them as excessive; rather, his scrupulosity pushed him to engage in plan-violating behaviors (namely, his compulsions). Additionally, this kind of excessive activity requires some degree of planning. After all, setting aside three hours to pray each day doesn’t just happen. Again, this suggests that the scrupulous individual is deficient at the level of planning management (or vigilance), not at the level of planning.
The real problem is that scrupulous individuals are inappropriately focused on a single commitment or care at the exclusion of all others. And this violates the complex pursuit of many different goals that require planning. Thus, their vigilance is sub-optimal in virtue of some lack of balance or efficiency in pursuing their goals. This suggests that there is an efficiency norm that uniquely governs vigilance.

Before specifying the efficiency norm, I first want to clarify the rationale behind the norm and consider why it applies uniquely to vigilance. As temporally extended social agents, we must aim for both intrapersonal and interpersonal coordination. When we make plans, we (and others) form expectations about our future actions. These expectations are part of what enables coordination both within and across agents. When we fail to live up to some norm of efficiency, we are not able to live up to the expectations that are set for us (whether personally or by others). This norm, however, does not apply to planning itself. Plans (or elements of a plan) do not automatically coordinate with other plans (or plan elements). Evidence for this comes from the fact that a plan can be differently organized while still being the same plan (consider: I can have plan elements of going to the grocery store, the post office, and the dry cleaner, but the order in which I execute the plan does not bear on the identity of the plan). This implies that the coordinating components of a plan are not part of the identity conditions of a plan. Thus, the essence of a particular plan is not tied (fundamentally) to the coordinating aspects of that plan; rather, the coordinating components are externally imposed.

So, when some agent formulates a complex plan that incorporates a number of subroutines, coordinating these elements of the complex plan are not built into the nature of the plan itself; rather, the function of vigilance is to achieve coordination between
different components of a plan (in virtue of its role in regulating planning). That is why the efficiency norm applies to vigilance alone, because the norm fits with the functional profile of vigilance.

In addition to this, we need to distinguish between two kinds of activities in which we regularly engage. Some activities are such that we aim to accomplish them while expending minimum effort. Take, for example, trips to the grocery store. The aim of going to the grocery store is to get whatever one needs as quickly as is reasonably possible (of course, one needs to maintain appearances, so you won’t go sprinting through the grocery store; conforming to social norms is something that constrains the speed at which one moves through the store). One tries to take the shortest route to the store, one aims to take the most efficient route through the aisles, and deliberation (if there is any) focuses on maximizing output while minimizing effort (assuming that this is a normal trip to the grocery store; if you’re the kind of person that enjoys strolling through the aisles, just substitute in a different activity, like going to the dentist, traveling to the airport, or taking a shower).

These activities seem importantly different from other kinds of activities that are closely tied to projects and commitments that we take to be valuable or meaningful. Consider, for instance, spending time with family, going on a date, or swimming (leisurely!) in the ocean (again, if these aren’t your cup of tea, just substitute in your preferred activities). These are the kinds of activities where it seems inappropriate to frame deliberation about these activities in terms of expending minimum effort; rather, it seems that we frame deliberation about these activities in terms of expending satisfactory effort, where satisfactory effort falls within a median range that crosses a minimum
threshold without exceeding an upper limit (an Aristotelian mean, as it were). An
important difference between those two activities is that in the case of minimum effort
activities, an agent ought to update or abandon her plan on the basis of new information if
that information conveys a more efficient way of realizing the plan, whereas in the case
of satisfactory effort activities, the agent should update or abandon her plan (or some
element of her plan) only if there is new information that continuing to engage in that
activity will begin to jeopardize the possibility of following through on other parts of
one’s plans (for example, you should enjoy your date, unless you become aware of
information that continuing on your date will jeopardize your ability to finish a project
that is due in the morning).

These considerations indicate that there are two different kinds of efficiency norm
that govern exercises of vigilance corresponding to the different activity types. We can
formulate the norms as follows:

*Vigilance Efficiency (Minimum Effort):* If Agent adopts or acquires some
complex plan $P$, at least some elements of which contain future-directed
intentions to engage in minimum effort activity, then Agent should be
vigilant with respect to considerations that allow Agent to maximize
realization of the minimum effort activity elements of $P$ while not
expending effort at a level that detracts from realization of other elements
of $P$.

*Vigilance Efficiency (Satisfactory Effort):* If Agent adopts or acquires
some complex plan $P$, at least some elements of which contain future-
directed intentions to engage in satisfactory effort activity, then with
respect to satisfactory effort activity elements of $P$ Agent should be
vigilant with respect to considerations relevant to the satisfactory effort
activities to a degree that does not preclude vigilance with respect to
considerations relevant to other satisfactory effort activities in $P$ or some
other complex plan that Agent has.
The efficiency norms then sets a standard for what counts as a relevant consideration for some agent relative to her overarching plans. We want agents to be vigilant with respect to relevant considerations in the appropriate circumstances. The function of the efficiency norms is to set the standard of relevance for a given set of considerations and establish a threshold of appropriate effort expenditure for a given context.

The efficiency norms seem to govern vigilance uniquely. No equivalent norm seems to constrain belief-, desire-, or intention-formation. As I suggested above, the reason why the efficiency norms uniquely govern vigilance is a function of the unique role that vigilance plays in our agential economy.

1.5 Conclusion

This chapter argues that planning agency introduces novel environmental pressures on the planning agent. In particular, planning agents face the Problem of Scarce Information. Vigilance solves this problem, and the various structural and functional considerations sketched here focus on what vigilance is such that it plays this kind of role in planning agents. Additionally, it seems that exercises of vigilance are governed by efficiency norms, and we evaluate vigilance itself relative to this norm.
2.1 Introduction

In 1964, the Indian government established the Central Vigilance Commission to address executive corruption. In 2017, the French National Assembly enacted new vigilance laws for multinational corporations operating in France. The law requires these corporations to practice due diligence in correcting human rights violations and environmental impacts throughout the production process. On January 28, 1852, Wendell Phillips, speaking to the Massachusetts Anti-Slavery Society, stated that: “Eternal vigilance is the price of liberty; power is ever stealing from the many to the few.”

While these uses of vigilance differ significantly from one another, they all appear to share a common core. Vigilance implies a certain kind of attentiveness for purposes of correction. It’s no wonder, then, that modern cognitive psychology tends to adopt a similar characterization of vigilance. Vigilance is described as “long-term attentive behavior” (Adams, 1965) or “sustained attention” (Stroh, 1971; see also Finomore, Matthews, Shaw, and Warm, 2009). In a review of vigilance research, Oken, Salinsky, and Elsas (2006) claim that: “Sustained attention to task is the most common scientific definition of ‘vigilance’. ” This definition has its roots in early vigilance research. Paul
Bakan & Manley (1963), for example, defines vigilance as a “form of sustained attention” in an early memo outlining research problems for the study of vigilance. Relatedly, other researchers define vigilance as a capacity for sustained attention rather than sustained attention itself (Parasuraman, Nestor, & Greenwood, 1989). A recent study defines vigilance as “the capacity to sustain attention to any environmental source of information over prolonged periods on watch” (Szalma, 2014: 1315). The consensus that emerges from these definitions is that vigilance is connected essentially to sustained attention.

The tasks used to study vigilance reflect this conception. These tasks require participants to respond on the occasion of detecting some infrequent cue or stimulus change over time. In order to successfully complete the task, the agent must sustain attention, even in the face of boredom and distraction. Task performance can be used to determine both the limits of attention and the factors that cause attention to deteriorate over time.

This line of thought fits well with the practical aims of vigilance research. Early on, researchers focused mainly on military applications, especially since vigilance research flourished during World War II and the nuclear panic of the Cold War (see McGrath, 1963: 3). In those times, psychologists and military personnel wanted to know how to maximize the efficiency of radar operators and individuals operating early warning detection systems. Over time, as these concerns faded and military equipment automated, the focus of vigilance research has shifted to workplace safety, designing work systems to minimize the probability of lapses of vigilance and to minimize the effects of such lapses.
Thinking of vigilance as a kind of attention or as a capacity to sustain attention informs the interventions designed to diminish (or eliminate) the likelihood of lapsed vigilance. For example, distractible schoolchildren are prescribed methylphenidate (commonly sold under the name Ritalin) as a treatment for hyperactive attention (NIMH, 2014). Methylphenidate functions through boosting catecholamine, an organic compound implicated in mediating the activity of attention (Clark & Nquoost, 2014). Since 1973, airlines require at least two pilots in the cockpit to protect against crashes related to inattentiveness (NTSB, 1973 & 1994). Hence, conceptual debates about the nature of vigilance have practical upshots.

In this chapter, I argue against the view that vigilance is either a kind of attention or a capacity to sustain attention. On my view, attending is a manifestation of vigilance. This reflects a more general claim about the nature of vigilance. Vigilance is a higher-order management capacity that directs lower-order psychological capacities and processes to facilitate acting over time. Attention is one capacity that vigilance directs, but it is not the only one. Vigilance also manages, among others, prospective memory and cognitive control (as discussed in Sections 2.3.2 and 2.3.3). Thus, on my view, the standard view of vigilance gets the relationship between attention and vigilance completely backwards.

Despite this implication, I deny that previous research on vigilance should be entirely disregarded. In fact, despite the historical roots of the sustained attention view of vigilance, other theories of vigilance offer an alternative characterization. Sir Henry Head, in one of the earliest papers on vigilance in modern psychology, described vigilance as “high-grade psychological efficiency” (1923: 132) or the general receptivity
of an organism to stimulation. Similarly, N.H. Mackworth (1957) defined vigilance as:
“…a state of readiness to detect and respond to certain specified small changes occurring at random time intervals in the environment.” While both Head and Mackworth’s conceptions of vigilance are overly narrow, their focus on efficiency and readiness is important. Defining vigilance as a general state of receptivity aligns with the view that vigilance enables organisms to effectively self-govern within dynamic environments. Empirical work on vigilance can, for the most part, be interpreted within this paradigm, and I suggest some reinterpretations of current empirical work toward the end of this paper.

2.2 Vigilance and Attention

Before diving into specifics about the relationship between vigilance and attention, it would help to have some general sense of the nature of vigilance and attention. This section offers an outline of both that, while incomplete, should provide the needed material for the rest of the argument.

Vigilance partially constitutes a global capacity for effective self-governance. In particular, vigilance is necessary for temporally extended plural goal pursuit. Human beings are planning creatures who aim to do many complex things simultaneously. On top of these complex aims, human beings are blessed with vastly limited brains that cannot actively manage lots of information simultaneously. So, human beings need a way to manage information in a way that supports plural goal pursuit while also respecting the limits of bounded rationality. Vigilance provides this management. In fact, vigilance functions as a kind of higher-order management system of lower-order psychological
capacities and processes. The whole set of plans and goals for some individual informs how their vigilance manages their psychology.

To keep the discussion general, let’s consider the following abstract example. The example will also help to identify three distinct operations of vigilance. Suppose that Santiago wants to X and to Y. These cannot be accomplished simultaneously, so Santiago decides first to X, then to Y. This requires that Santiago act over time and, so, requires that Santiago be vigilant. This entails three things. First, Santiago must monitor for the right opportunity to begin to X and, upon the occasion of attempting to X, he must also determine whether his performance is coming up to relevant standards of X-ing. Monitoring, then, includes an awareness of when to begin acting as well as an ongoing assessment of one’s acting. This account of monitoring implies two further operations of vigilance. When Santiago detects the opportunity to X, he must recall the intention from memory and implement the intention to X. However, when Santiago is monitoring his current X-ing and detects an error, he must implement relevant changes to adjust his X-ing or stop X-ing altogether. Monitoring alone, however, does not issue in adjustment. Santiago also regulates his X-ing. Monitoring and regulating are closely linked. When Santiago monitors current performance, this monitoring informs regulative components that determine whether to adjust or abandon current activity. So, error detection likely initiates a decision process about whether to give up or how to adjust one’s behavior. Santiago might also regulate in switching from X-ing to Y-ing. That is, the occasion for Y-ing might unexpectedly present itself, so that Santiago should switch tasks and revise his plan. Switching, here, would also require implementing a new intention. However, this
implementation does not eliminate the need to \( X \). Santiago stores the intention to \( X \) in memory, and monitors for further opportunities to take up \( X \)-ing again.

This example identifies three operations of vigilance: monitoring, implementing, and regulating. Both of these operations have two distinctive aspects, what we might call internal and external aspects. The external aspect of monitoring aims to detect opportunities to act or switch tasks, while the internal aspect of monitoring aims to detect errors in action. The external aspect of regulating aims to determine whether to switch from one task to another on some occasion, while the internal aspect of regulating aims to determine whether to adjust or abandon one’s current behavior. The aspects of implementation do not divide into internal and external, but they do divide into three distinct categories: implement an intention, switch from one intention to another, and remove an intention. The agent’s overall goals and plans govern how these operations are carried out. This reflects the fact that vigilance aims to manage the limited resources of an agent’s psychology in light of pursuing complex, temporally extended goals.

Attention is more difficult to characterize, even generally. There is no widespread consensus about the nature, function, or even phenomenology of attention. Hence, I will avoid offering any necessary or sufficient conditions on attending to avoid inviting controversy. Instead, I will focus on the factors that initiate and guide attention over time. These factors are often used to distinguish different kinds of attention. My argument is that, for any possible kind of attention, instances of that kind manifest vigilance. So, whatever attention ends up being, instances of attention manifest vigilance.

Finally, the notion of manifestation deserves comment. A manifestation of some disposition is an event the occurrence of which counts as a realization of the disposition
(see Audi, 1994: 421). In the same way that dissolving counts as a realization of solubility, instantiating a particular pattern of attending counts as a realization of vigilance. These realization facts imply the truth of some counterfactual conditionals, such as ‘If an agent is vigilant to some degree, then the agent attends in some particular way’. Part of the argument that follows, then, aims to show that no particular instantiation of attending is not also a realization of an agent’s vigilance. This, however, does not entail that every realization of vigilance is some instance of attending.

2.3 Attention Manifests Vigilance

In this section, I defend the claim that instances of attending always manifest vigilance. The standard view characterizes vigilance as a mode of attention or as a capacity to sustain attention. If my argument is correct, the standard view is wrong. Attention stands to vigilance as a manifestation stands to a disposition.

The previous section mentioned that discussions of attention are often fragmented and messy. Part of what makes the literature on attention messy is that people countenance many different kinds of attention. Wayne Wu (2014: 31) has identified five different contrasts used to classify attention: top-down/bottom-up; endogenous/exogenous; goal-directed/stimulus-driven; controlled/automatic, and; voluntary/involuntary. These contrasts divide into two different categories. One category

---

9 This does not imply that dispositions of any kind can be fully analyzed in terms of corresponding conditional statements. I acknowledge all of the problems with the counterfactual analysis of dispositions. What I am committed to here is that, generally, when an object has some dispositional property $D$, the object will typically manifest $D$ in the right circumstances. That’s not meant to be a particularly illuminating statement on the nature of dispositions, but I’m happy to help myself to the notion of dispositions that almost everyone else uses and leave the technical details for others.
concerns how attention gets initiated, while the other category concerns how attention is
directed over time (Wu, 2014: 34). The first category maps, roughly, onto the distinction
between top-down and bottom-up attention, where top-down attention encompasses any
instance of attending initiated by goal-directed states and bottom-up attention
encompasses any instance of attending that is not so initiated (hence, bottom-up and top-
down attention are exhaustive, mutually exclusive categories of attention). The second
category maps, roughly, onto the distinction between controlled and automatic attention,
where controlled attention encompasses any instance of attending that is guided by an
explicit task representation. Automatic attention encompasses all other instances of
attending that are not controlled.

The distinction between bottom-up and top-down attention offers an exhaustive
categorization of the causal antecedents of attending, while the distinction between
controlled and automatic attention exhaustively characterizes the factors that guide
attending over time. In what follows, I will discuss only these distinctions. I argue that
top-down, bottom-up, controlled, and automatic attention all manifest vigilance. Insofar
as these distinctions exhaustively categorize attention, all attention manifests vigilance. I
begin with top-down attention, then go on to discuss controlled, automatic, and bottom-
up attention.

2.3.1 Top-Down Attention

The kind of state that can initiate top-down attending (or attending that counts as
an instance of top-down attending) is a goal-directed state. Adopting a goal-directed state
entails adopting commitments to guide attention in ways that support realizing the object
of the goal-directed state. These commitments, in turn, structure vigilance such that subsequent exercises of vigilance are assessed in light of standards set by those commitments. When these commitments initiate attention, the subsequent attending manifests the agent’s vigilance (likewise, when these commitments fail to initiate attention and the agent fails to attend, this also manifests the agent’s vigilance, though it manifests suboptimal vigilance).

Someone might wonder why positing vigilance is necessary, when it looks like either the goal-directed states or the commitments to guide attention do the relevant work of initiating attention. Why say that attention manifests vigilance rather than those underlying states? Adopting a goal-directed state need not issue immediately in attending. For instance, suppose you want to stop at the library to pick up a book. It’s not convenient for you to do it right now, so you plan to do it at the end of the day on your walk back to the car. Adopting the goal-directed state of going to the library does not immediately issue in attending that supports going to the library. So, the goal-directed state itself is not sufficient to initiate attending.

When you plan to go to the library at the end of the day, there’s no need to think about it all day. You store the intention in memory and recall it only at the appropriate time. You might use some cue to trigger recall of the intention. Perhaps it’s the clock striking 5 or the sight of the library building itself. Whatever the cue is, perception of the cue is insufficient. In order for the cue to trigger recall, something must classify the perceptual stimulus in the right way. There must be some recognitional capacity tuned to the agent’s goals and plans that functions to properly classify perceptual stimuli and identify relevant cues (consider the fact that, on a different plan, seeing the library might
trigger recall of a different cognitive state; the perceptual stimulus itself is not rich enough to function as a cue on its own). This recognitional capacity identifies the cue and implements the appropriate goal-directed state. These functions appear to map onto the functional profile of vigilance. Because of this, we have reason to presume that the recognitional capacity is vigilance.

An exercise of vigilance, on some occasions, issues in recall of some goal-directed state while also initiating attention to support realizing the aims of these states. The connection between goal-directed states and vigilance implies a connection between vigilance and top-down attention. Every instance of attending that counts as top-down attention manifests vigilance in virtue of the fact that the implementation of a goal-directed state subsequently initiates attention in ways that support goal-relevant behavior.

2.3.2 Controlled Attention

Top-down attention is distinct from controlled attention. Briefly, what makes some attending an instance of top-down attention is whether a goal-directed state initiates the attending. What makes some attending an instance of controlled attention is a matter of how the attending is guided over time. Thus, instances of top-down attention need not also be instances of controlled attention.

Given the relationship between controlled attention and guidance, we need to determine what guidance consists in before we can understand the relationship between controlled attention and vigilance. The following schema captures the notion of guidance I’m interested in:

\[ X \text{ guides } Y \text{ toward being in state } Z \text{ only when:} \]
(a) $X$ monitors the status of $Y$ (relative to its approaching being in state $Z$)

(b) $X$ alters $Y$ when it deviates from moving toward $Z$-state

This notion of guidance suggests that the factors that make attending count as an instance of controlled attention are those that monitor and regulate attention over time. What, then, monitors and regulates attention?

To answer this question, I suggest we take a step back to ask why attention must be guided and regulated at all. Some tasks we perform are complicated and consist of various subtasks that must be performed in a specific order. Many of these tasks consist of numerous subtasks such that the entire task cannot be explicitly represented all at once. To perform these tasks, we keep in mind only the current subtask and bring to mind the next subtask only at the appropriate time.

As noted above, when an individual aims to perform some task, that individual also adopts commitments to guide attention in ways that support task performance. Task complexity dictates the need to guide attention over time to the currently relevant aspects of task performance. Thus, the complicated nature of the tasks we sometimes performs is part of the reason why attention requires guidance.

Again, one might wonder whether the goal-directed state itself might guide attention. Consider someone who is performing the task of tiling their bathroom. This is a complex task that requires measuring, cutting, placing, and grouting the tile, with each tile taking several minutes to prepare. Each stage of the task requires focus, especially because small errors iterated over the whole bathroom floor can disrupt the entire tile field. Beyond just having a goal-directed state represent task features, individuals need a recognitional capacity that detects the need to transition between sub-tasks and guide
attention to the features of the task representation that are currently relevant. The functional profile of the recognitional capacity, here, gives us good reason to think that the recognitional capacity is vigilance.

This is similar to the role ascribed to vigilance above in discussing how attention gets initiated. But here, vigilance goes beyond initiating attention and directs attention over time to relevant features of a task representation. There is also a regulatory role for vigilance, as well. Consider, again, the person tiling the bathroom (let’s say the person’s name is Joel). Joel’s measured out the tile and is now taking it to the garage to cut it with the masonry saw. However, in the garage, Joel’s eye catches the golf clubs in the corner of the garage. Joel had been planning for some time to clean his clubs, and with the weather warming up he knows that it needs doing sooner rather than later. Now, suppose that several minutes later Joel catches himself cleaning the clubs instead of cutting the tiles. Something needs to pull Joel’s attention from the clubs back to cutting the tiles. This pulling might consist in recognizing that the current moment is not the best when considering trade-offs between a Clean Now/Cut Later plan and the Cut Now/Clean Later plan. This kind of trade-off reasoning fits with the regulatory dimension of vigilance. So, again, it seems we have good reason to suppose that vigilance both monitors and regulates attention over time. Hence, any instance of attending that counts as controlled attending manifests vigilance.

It makes sense that vigilance both initiates and guides attention. Vigilance, recall, solves a problem of planning in creatures with scarce cognitive resources. Human beings form plans to streamline deliberation, but these plans are either too complicated or too far off in the future to hold in mind. So, we store plans in memory. Vigilance aims to call
these plans to mind at the right time. This recall directs attention to a task only when focusing on the task is appropriate (relative to all of the agent’s other goals and commitments). Similarly, vigilance guides attention so that when the task is done, attention is directed. Alternatively, if environmental or situational changes occur, attention can be modified accordingly. In both the initiation and guidance phase, vigilance manages attention according to the interests of the individual. This also implies a feedback loop between attention, memory, and vigilance. When vigilance initiates attention, there is also monitoring. This monitoring informs how vigilance integrates with memory to bring online new tasks or task subroutines based on current performance.

2.3.3 Automatic Attention

There seem to be kinds of attention that are disconnected from tasks. Instances of automatic and bottom-up attention, for example, are defined in terms of their lacking goal-directed initiators and guidance over time. Attending counts as an instance of automatic attention if there is no goal-directed state that guides the attending over time; instead, perceptual states or habits guide attention over time. Consider, for example, driving home from work. Assuming that you’re not fresh off moving to a new country or a new career, driving home is a routine activity. Your attending to the road might not be governed by any particular goal-directed states. You might be focusing on something that happened earlier in the office or on a presentation you’re giving next week. But this doesn’t prevent you from attending to changing traffic lights, merging, hitting the turn signal, or whatever else you need to do to get home. Habits and perceptual states regulate this form of attention. These states tend to be more rigid, so the resulting attention exhibits less flexibility than other forms of attention. In concrete terms, this means that
instances of automatic attention respond well to green lights switching to red and normal merging behavior. If, however, one intersection light is flashing red, that probably generates a switch to more regulated forms of attention, because it’s rarely advisable to pass off handling the unknown to an inflexible system. But, when operating within the bounds of the ordinary, automatic attention has a reliable track record.\textsuperscript{10} You rarely drive off the road when coming home from work, even if you spend most of the time thinking about things other than driving home.

Automatic attention is typically good enough for familiar tasks. However, someone might wonder why we utilize automatic attention at all given its inflexibility, especially when we have a more reliable form of attention at our disposal. Controlled attention, however, is computationally expensive. Recall from Section 2.2 that human beings have limited cognitive resources. This includes a small amount of working memory space. Insofar as controlled attention requires working memory to actively maintain explicit task representations to guide attention, utilizing controlled attention comes at the expense of being able to deploy such attention elsewhere (for the relationship between controlled attention and working memory, see Awh, Vogel, and Oh, 2006; Kane, Bleckley, and Conway, 2001; Engle & Kane, 2004). This means that, while controlled attention is more reliable, it also occupies most of an individual’s available cognitive resources. Automatic attention, however, can be used less discriminately. This

\textsuperscript{10} This is a controversial point, as there is some evidence for flexible automatic processes (see Hassin, Bargh, and Zimerman, 2009). However, the claim here is not that automatic processes are completely inflexible. As noted above, within relatively well-defined task domains, automatic processes can exhibit flexibility (differentially responding, say, to different colored lights at the intersection). The claim here is that automatic processes exhibit vastly less flexibility than other kinds of processes that goal-directed states initiate and regulate.
creates a trade-off that human beings must navigate in determining when to use automatic or controlled attention.

Some scenarios require selecting whether to engage a task with automatic or controlled attention. It might be that, given a history of success at performing the task, one may allocate automatic attention to a task in order to focus elsewhere. The example of driving home from work illustrates this point. You don’t need to allocate controlled attention to driving because you’ve done it so many times. Utilizing automatic attention allows one to focus on what to buy one’s fiancée for Christmas this year (actually, by the time I finish this I’ll probably be thinking of what to buy my wife for Christmas; but that’s my problem, not yours).

Hence, when an agent attends to some task such that the attending counts as an instance of automatic attention, this reflects implementing a strategy regarding cognitive resource allocation. The issue of whether and where to deploy one’s attention reflects higher-order cognitive management policies and the agent’s underlying evaluations of expected task demands balanced against expected probability of success. Roughly, these calculations aim to determine where to distribute psychological resources over time in a way that facilitates plural goal pursuit. Given that human beings can only do a small number of things at any one time, devoting resources to one activity has to be balanced against all of the other currently available activities and whether pursuing one activity puts one in a good position to pursue further activities at a later point. If some individual expects a high degree of success for some activity even with a small amount of resources, then it doesn’t make sense to devote many psychological resources to performing that activity. Thus, highly routinized activities (that occur in familiar task domains) do not
require the same degree of psychological resources as, say, novel activities or activities in unfamiliar domains.

This description aligns with how human beings actually manage their psychological resources over time. Recent studies show that human beings strategically shift between different kinds of attending based on various factors related to expected task demands, environmental regularity, and expected value of control. For example, individuals tend to use reactive (bottom-up) forms of attention in familiar environments more than proactive forms of control (Lewis-Peacock, Cohen, and Norman, 2016). Individuals tend to mind wander at regular intervals when performing low-demand tasks in stable environments (Seli et al., 2018-a). Tasks that require more complex planning are associated with use of model-based control rather than model-free control, the former of which is associated with controlled, top-down attention (Kool, Gerhsman, and Cushman, 2018). All of these show that when someone is utilizing non-controlled processes to perform tasks, this reflects an underlying evaluation of the costs and benefits associated with allocating psychological resources in particular ways over time. These evaluations themselves reflect the regulative dimension of the activity of vigilance.

Attending that counts as an instance of automatic attention are manifestations of vigilance precisely because agents always face dilemmas about how they manage trade-offs in utilizing cognitive resources. Automatic attending, as it were, reflects a cognitive strategy governed by the agent’s vigilance.

2.3.4 Bottom-Up Attention

The final kind of attention to consider is bottom-up attention. This kind of attention is observed in cases of attentional capture, where an intense or unexpected
stimulus grabs attention without the agent’s intending to shift attention (e.g., hearing a loud bang in the restaurant). Like automatic attention, bottom-up attention seems disconnected from tasks and goal-directed behavior. It exhibits many of the characteristics of a reflex and, thus, might seem unlikely to bear any important connection to vigilance.

This, however, ignores certain interesting features of bottom-up attention. Attentional capture responses scale in proportion to an organism’s familiarity with the environment (Grillon, 2002; Grupe & Nitschke, 2013). This provides some evidence that bottom-up attention is sensitive to an organism’s higher-order states of awareness. Following on this, attentional capture responses may reflect policies to investigate potential threats, particularly in informationally translucent environments (this, in turn, explains why bottom-up attention is mitigated in proportion to the informational transparency of the environment). These policies need not be intentionally adopted. Policies concerning threat detection may reflect adaptive habits shared by any creature with vigilance because the benefits of having a sensitive, situationally fluid threat detection mechanism vastly outweigh the costs. However, while these policies are likely not intentionally adopted, they can be intentionally calibrated. Top-down regulation can modulate the response profile of bottom-up attention (Wells & Matthews, 1994; Cisler & Koster, 2010).

As was the case with automatic attention, bottom-up attention responses appear to reflect strategic allocations of cognitive resources that aim to manage an individual’s psychological economy. For example, attending that counts as an instance of bottom-up attention may reflect a strategy to respond to potential threats in relatively unfamiliar
environments. So, even though non-goal-directed states initiate attending that counts as bottom-up attention, the fact that these non-goal-directed states are salient at all reflects the agent’s underlying vigilance.

There are two ways to classify attention: its causes and the internal dynamics of attending. This implies that there are four different kinds of attention, top-down/bottom-up and controlled/automatic. For all of these, vigilance is a crucial antecedent. Either vigilance plays a role in initiating attention or in guiding attention over time. Given this relationship between vigilance and attention, we have good reason to believe that attention manifests vigilance.

2.4 Performance Breakdowns

The previous section shows that attention manifests vigilance. But does attention exhaust the constitutive manifestations of vigilance? In other words, is vigilance simply a disposition to attend? The characterization of vigilance offered in Section 2.2 suggests not. Vigilance manages an individual’s psychological resources, and there are more psychological resources available to a planning agent than just attention. So, even though attention manifests vigilance, not every manifestation of vigilance is an attending. To show this, I identify a number of failures of vigilance that do not seem to be failures of attention. From Section 2.2, we have some sense of the functional and structural features of vigilance that should help guide identifying certain occurrences as instances of vigilance (or failures of vigilance). Difficulty arises, however, in the case of attention, where there is no widespread agreement about the structural, functional, or phenomenological properties of instances of attention. This forces us to adopt a
somewhat clumsy methodology. After reviewing certain instances of failures of vigilance, we will have to run through a catalogue of various views on the nature of attention to see whether any particular view can explain failures of vigilance as failures of attention. As we will see, no single view of attention seems well positioned to explain every failure of vigilance as a failure of attention.

In the remainder of this section, I outline a number of performance breakdowns (i.e., failures of vigilance). In Sections 2.4.2 & 2.4.3, I consider whether we can explain these performance breakdowns as either prospective memory errors or mechanical/action failures, respectively. I argue that we cannot. This sets up the discussion in Section 2.5 about whether we can explain these performance breakdowns as failures of attention. In Section 2.6, I claim that reflecting on general features of performance breakdowns sheds light on one of the functional properties of vigilance.

2.4.1 Examples of Performance Breakdowns

I present four cases that fall into two categories: 1) Failure to Carry Over (failure to recall a task set & failure to preserve a task set), and; 2) Failure of Foresight (double booking & coordination failure). Collectively, I refer to these as performance breakdowns.

First, let’s characterize instances of failure to carry over, beginning with failure to recall a task set. Forgetful cooking, per the National Fire Protection Association, causes almost half of all home fires in the United States (Ahrens, 2016). This is where people forget that they failed to turn the stove off before going to do something else. People usually remember the stove when their kitchen is already on fire. These appear to be cases where some agent fails to recall a task set that they (at an earlier time) planned to
execute. For instance, someone might plan to put the pasta in the water when it’s ready and go do something else while the water’s heating. Somewhere between planning and burning, the agent succumbs to a failure of recall. Unattended cooking is certainly dramatic, but it should feel familiar enough. All of us have, at some point, forgotten to pick up a key ingredient at the grocery store or forgotten to send a response email. These are all examples of failure to recall a task set (or, more precisely, these are candidates for being failures to recall a task set; of course, other factors might cause these sorts of mistakes to occur). The agent plans to perform a task in order to realize some goal at a later time, but she fails to recall that task at an appropriate time.

Unattended cooking is slightly different than another kind of performance breakdown. Consider the following true story (names have been changed):

*Hot Car.* Kate normally dropped her daughter Sam off at daycare before working from home. But one summer morning, Kate had an early meeting and her daughter was sleeping in. This meant Brad—Kate’s husband—had to take Sam in before heading to work. Unfortunately, Brad forgot to drop his daughter Sam off. By the time he remembered, the little girl had been sitting inside a car, windows closed, on a summer day in Texas for several hours. EMT’s declared Sam dead a mere 80 minutes after Brad called 9-1-1.\footnote{This is based on a true story (Pelletiere, 2016).}

This failure is less common than forgetful cooking. Still, 39 children died of heatstroke in 2017 because of parental forgetfulness (Null, 2018). Hot Car Syndrome (a term coined by Weingarten, 2009) does not look like failure to recall a task set; rather, HCS seems to be the result of failing to preserve a task set until successful action occurs. In failure to recall a task set, the agent plans to perform some task at a later time, and stores the task
representation in memory. In failures to preserve, the agent does not store the relevant
task. She maintains the task, but at some point that task is replaced with something else—
usually, at least in cases of HCS, the task is some entrenched habitual routine where a
parent normally breaks routine to take their child somewhere before work (as in the story
of Brad and Kate). In the car, the parent forgets and comes to perform a habitual task
(driving to work) without noticing the switch (see Amaya, Manuscript).

The common core element of failures to carry over is that the agent fails to
apprehend that a presently available consideration is relevant given some previously
adopted plan. ‘Presently available’, here, is meant to signal that the agent possesses all
the right sort of information to avoid a performance breakdown. The parent certainly has
the information that the child is in the backseat, that he did not go to the daycare, etc. The
forgetful cook knows that you should not leave the stove running for too long without
checking (this does not mean that the agent has the information in mind, but the
information is still easily accessible to the agent without requiring interaction with the
environment). Thus, despite the differences between these two kinds of performance
breakdowns, there is similarity at the level of what kind of failure the breakdown exhibits,
namely failure to apprehend the features of a present context that are relevant in light of
one’s plans.

Failures of foresight constitute the second major category of performance
breakdowns. A recent example of this occurred in Tulsa, Oklahoma. Alicia Skaggs and
Austin Swartz wanted to get married in the Oklahoma Jazz Hall of Fame. However, the
day before the wedding, a guest informed the couple that the Hall was hosting a different
event on the day of the wedding. After some scrambling, the couple moved the ceremony
to an alternative venue. The manager of the Hall, Jason McIntosh, never explained what happened and failed to inform the couple—despite meeting with them three days before the wedding—that the venue was double booked. The couple is currently considering legal action against the Hall and its management (Krehbiel, 2017). What makes double booking a performance breakdown is that the agent commits to a task set that is incompatible with some task set that the agent previously planned to perform (and committed to performing). Thus, Jason McIntosh commits to acting in such a way that is practically impossible. More mundane examples of double booking include committing to go to lunch with a friend despite the fact that you previously planned to meet with a student at the exact same time or accepting a conference invitation that conflicts with family vacations or holidays.

The final sort of performance breakdown is similar to double booking. Consider the following example. A friend asks you to write a paper for a conference or volume. The topic is one that you have thought about for quite a bit, but you have no manuscript ready to go. The deadline for submitting the paper is near the end of the term (assume, contrary to present practice, that this is a hard deadline; the request is made, let’s say, to fill in for a contributor that dropped out at the last minute, so your friend needs the paper at the deadline). Suppose that you commit to writing the paper, despite knowing full well that: (1) you will have to write a manuscript essentially from scratch, and; (2) you will be busy with teaching and administrative duties at the end of the term. Despite committing to writing the paper, you also must drop out at the last minute due to increased workload and an inability to put enough time into the manuscript. In this case, you do not commit to task sets that are incompatible (as in the case of double booking); rather, you commit...
to a task set that you likely will not remain committed to (and for which you have good evidence that you will not follow through).

The core element in cases of failures of foresight is a failure to coordinate one’s present and future selves (hence, I’ll refer to these as coordination failures). In other words, there is a failure to apprehend that some future consideration is relevant to one’s current planning and task set acquisition. This marks the crucial difference between failures to carry over and failures of foresight. In the former case, the agent fails to apprehend that a present consideration is relevant given prior planning, whereas in the latter case, the agent fails to apprehend that a future consideration is relevant to one’s current planning.

Despite the differences between these performance breakdowns, there seems to be at least one significant unifying factor. All involve a failure to move from an adopted plan to relevant features of an action context. The cases divide as to whether the action context is in the present or the future (and, perhaps, further divisions could be derived by dividing time-based features of contexts and space-based features of contexts). But, at a high level of generality, these contextual factors are secondary to the main issue, namely that performance breakdowns are a failure to act on one’s plans as a result of failing to apprehend some relevant feature of an action context.

2.4.2 Are Performance Breakdowns Memory Failures?

Before considering whether these performance breakdowns are best construed as failures of attention, we need to consider whether we can explain these phenomena in terms of other psychological constructs. For instance, why not think that all of these performance breakdowns are just prospective memory failures?
This proposal has some intuitive merit. Failure to recall a task set, naturally, seems like a failure to recall a task representation stored in memory, whereas double booking and coordination failures also seem to qualify as failures of recall. Here, it seems that we can explain each failure as a failure of prospective memory without appealing to anything like vigilance. Notice, however, that failure to preserve a task set does not seem to be a prospective memory failure, especially if you think that task set maintenance is not part of the function of prospective memory (it might be part of the function of working memory, but working memory and prospective memory are distinct constructs; for discussion of task set inertia and working memory, see O’Reilly, 2006). This might not seem like a problem, though, since we could explain some performance breakdowns as failures of prospective memory and others as failures of attention. There are two reasons for skepticism about this strategy. First, as I argue in the next section, none of these performance breakdowns seems to qualify as a failure of attention. Second, there are reasons to think that some of these performance breakdowns are not fundamentally prospective memory failures, even though a prospective memory failure might partially constitute the agent’s omission.

Cognitive scientists have recently distinguished between two forms of control implicated in prospective memory tasks: proactive and reactive (see Braver, 2012). Proactive control relies on storing task representations in working memory and maintaining these representations continuously up until the moment of action. Reactive control relies on storing task representations in long-term memory and associating a cue with the task representation. This form of control utilizes bottom-up attentional capture mechanisms that detect the cue that then triggers recall of the task representation (see
Cohen & O’Reilly, 1996; Botvinick et al., 2001). Proactive control is highly reliable, but computationally expensive, as we can hold only a limited number of representations in working memory at any given time (this way of putting things makes it seem like the cost of proactive control derives from a structural limitation of working memory; however, these structural limitations might ultimately bottom out in utilizing efficient computational principles, as Cohen, 2017 argues). Reactive control is less computationally expensive (as the storage capacity of long-term memory is quite large), though less reliable given the reliance on attentional capture and the easily disrupted associative links between cues and task representations. Hence, cognitive efficiency requires balancing use of proactive and reactive control (Boureau et al., 2015; Kool, Shenhav, & Botvinick, 2017; Shenhav, Botvinick, & Cohen, 2013).

Recent evidence also suggests that individuals can strategically select to engage a particular form of control based on anticipated task demands (Lewis-Peacock, Cohen, and Norman, 2016). This indicates why the aforementioned action failures are not always fundamentally prospective memory errors. When we focus on the end stages of performance breakdowns, we are inclined to see these as prospective memory failures. But these failures are predicated on earlier failures of improper strategy selection in memory encoding. This strategy selection is not part of the function of memory itself (at least not prospective memory, which is typically mapped to associative mechanisms that bind cues to representations and recall mechanisms that activate task representations). In a word, then, these performance breakdowns are, in some instances, failures of prospective memory management, not failures of prospective memory proper.
It might turn out that some performance breakdowns result from mere memory failures. But this does not affect the argument. The point here is that there is some way of carving up different kinds of performance breakdowns such that one category of such breakdowns cannot be explained fundamentally as mere memory failures. The cases mentioned in the previous section seem to fall into just such a category.

2.4.3 Are Performance Breakdowns Mechanical or Action Failures?

One final bit of bookkeeping is necessary before moving on to attention. Might performance breakdowns just be mechanical failures or action failures? Why not think that these represent a kind of ‘glitch’ in the cognitive life of the agent?

These performance breakdowns do not seem to be mechanical. There’s nothing that seems to interfere with the agent’s cognitive capacities (i.e., there’s no noise that seems to interfere with information processing and nothing that disrupts functional connectivity). Similarly, these do not seem to be cases of action failure. Action failure seems closest to accidental behavior or clumsiness. But notice that, in performance breakdowns, the agent acts fluidly and competently despite acting contrary to her plans. Performance breakdowns, in this way, differ significantly from, say, dropping your phone accidentally. Likewise, the failures that constitute performance breakdowns do not seem to be wholly subpersonal in the way that mechanical failures are. Performance breakdowns exhibit rational patterns, and they do not have the external appearance of clumsy or accidental behavior. From the outside, a performance breakdown looks like a perfectly ordinary case of intentional action.

Of course, you can imagine someone saying that it’s perfectly consistent to construe the aforementioned performance breakdowns as mechanical or action failures.
And that’s correct. But we might also construe these performance breakdowns as something else without any significant leaps of the imagination. Those are the performance breakdowns that are of interest here. The argument here can tolerate some performance breakdowns being mechanical failures. But not all of them are, and that’s enough for purposes of the present argument. Suffice it to say, while we might be able to explain some performance breakdowns as merely mechanical or action failures, surely this will not work as a general strategy.

2.5 Attention and Performance Breakdowns

In this section, I consider different theories of attention to see whether any of them can offer a general explanation of performance breakdowns in terms of failures of attention. In particular, I focus on cognitive unison theories of attention (Mole, 2011), structuring theories of attention (Broadbent, 1958; Watzl, 2017), broadcast theories of attention (Smithies, 2011; Prinz, 2012), and selection for action theories of attention (Wu, 2014). For purposes of space, I will not go into the theoretical or empirical motivations that support each view; instead, I will argue that none of these theories offers a general explanation of performance breakdowns. In conclusion, I consider some reasons why failures of attention, in general, might not provide a general basis for explaining performance breakdowns.

2.5.1 The Cognitive Unison Theory of Attention

Christopher Mole (2011) defends a theory of attention as cognitive unison. For Mole, attention is neither identical to a particular cognitive process nor functionally
localized; instead, attention is a mode of acting such that one’s activity manifests cognitive unison. Mole describes cognitive unison in this way.

Let A be an agent, T be some task that the agent is performing, and call the set of resources that A can, with understanding, bring to bear in the service of T T’s ‘background set’. Then, A’s performance of T displays cognitive unison if and only if the resources in T’s background set are not occupied with activity that is not in the service of T (Mole, 2011: 51).

Hence, when all of an agent’s relevant understanding is brought to bear on task performance, then the agent is, on Mole’s view, performing the task attentively.

It is true that in performance breakdowns, an agent fails to manifest cognitive unison. But does this make the performance breakdown a failure of attention (on Mole’s view)? No. Instead, the performance breakdown is a failure to appropriately direct attention. In Mole’s terminology, the performance breakdown is a failure to bring certain available task-relevant resources to bear on one’s task performance. But it is not part of the function of attention to marshal these resources. For Mole, successful unison displays acting attentively.

Therefore, performance breakdowns reliably co-occur with failures of unison, but failures of unison do not ultimately explain the performance breakdown because the failure of attention to fulfill its functional role does not explain why the performance breakdown occurs. For this reason, the unison view does not seem to offer a satisfactory account of failures of attention in relation to performance breakdowns.

2.5.2 Structuring Theories of Attention

Structuring theories of attention enjoyed a measure of popularity in mid-20th century cognitive psychology. Donald Broadbent developed a highly influential theory of
attention where the function of attention is to serve as a mechanism for managing
capacity limits in information processing channels (Broadbent, 1958). This led to the
idea that attention functions as a bottleneck between lower- and higher-order processing
channels. The bottleneck metaphor resulted in the debate about where in the perceptual
processing hierarchy the attentional bottleneck occurs (the early vs. late selection debate).
Most now consider that debate wrongheaded for several reasons. First, the empirical
predictions of the view have not been confirmed (see O’Connor et al., 2002). Second, the
framework incorrectly assumes that perceptual processing is linear, rather than dynamic
(Hommel et al., 2001). Finally, recent evidence suggests that the occurrence of
bottlenecks in information processing might occur in different locations given differential
task demands (Lavie, 2005).

A more sophisticated theory of attention within the Broadbent tradition takes
attention to be a general-purpose, structured, information integration process (Watzl,
2017). This shares with Broadbent’s view the idea that attention structures the flow of
information, but it denies that the structuring activity is localized somewhere in the
perceptual processing hierarchy. On Watzl’s view, attention is constituted by a structure
of the mind, where this structure consists in priority relations between mental states
(2017: 67-68). Deployments of attention are governed by these priority structures, and
these prioritization relations are partially constitutive of acting attentively (2017: 70).

12 Mole (2017) points out that Broadbent rarely uses the term ‘attention’ and did not claim to be
offering a theory of attention. However, this might be due to the fact that cognitive psychologists were
hesitant to invoke the concept of attention given the outsized influence of behaviorism in psychology in
Broadbent’s day.
Thus, for Watzl, the function of attention is to integrate and process information that is relevant relative to one’s priority structures.

This sort of view seems to explain performance breakdowns, especially since all of the examples describe an agent failing to integrate plan-relevant information in a way that guides subsequent action. However, one problem is that attention emerges from priority relations among mental states, and we can ask what determines these priority relations. Some performance breakdowns might result from failed prioritization, which would then explain the mere co-occurrence of performance breakdowns and failures of attention.

In this context, that is not a fully satisfying response. For instance, consider the Hot Car example of a failure to recall a task set. It seems highly implausible to say, in that case, that the parent fails to appropriately prioritize the child over getting to work; instead, the more natural interpretation of the case is that the parent has the correct prioritization, but there’s a failure to process and integrate relevant information. Hence, the performance breakdown results from a failure of attention to fulfill its function.

But we need to be careful here. In performance breakdowns, one of two things might be going on. Either there is a change in the agent’s priority structures or there is not. If there is a change in priority structure, then the fundamental question to ask is what caused the change in priority structure. Since attention does not play a role in that, but rather emerges from priority structures, failures of attention cannot explain performance breakdowns that result from changes in priority structures.

If there is no change in priority structure, we still cannot definitively tie the performance breakdown to a failure of attention, because Watzl’s functional
characterization of attention is both mechanically and temporally ambiguous. The mechanical ambiguity stems from not knowing how attention interacts with psychological mechanisms that mediate between priority structures and deployments of attention. Consider, for example, the Hot Car case. What might be happening here is that the parent has certain priorities (drop the kid off, get to work) that inform the formation of goals and the acquisition of relevant task representations. Crucially, however, these representations may be cognitively underspecified (see Reason, 1990; Sellen & Norman, 1992), which leaves the agent susceptible to habitual substitution (roughly, when task representations are underspecified, the agent is prone to saturate these representations with habit-based routines). This seems to describe the Hot Car case. The agent commits to an underspecified task set, and at some point habit-based routines are substituted for the goal-relevant routines. A process of this sort does not require a change in priority structure; instead, the failure occurs somewhere between prioritization and deployments of attention.

The temporal ambiguity concerns the priority structure against which we measure the successful deployment of attention. The Hot Car case does not bring out this ambiguity, so consider a case of goal shift in response to temptation (cf. Holton, 2009). Sometimes, we commit to courses of action that we know will be difficult to follow through on (e.g., training for a marathon) or that we will want to give up on at certain points (e.g., quitting smoking). Suppose I resolve to quit smoking, which provides me with a certain kind of mental priority structure. Formally, my resolution at \( t_1 \) confers priority structure \( P \), which deploys attention toward \( A \). At some later point, when I undergo a craving, I abandon my resolution and enjoy a cigarette. Thus, at \( t_2 \), I abandon
my resolution, which confers priority structure S (for smoking) and deploys attention toward B. Abandoning the resolution is a kind of performance breakdown, as it’s a failure to coordinate (in the example from the previous section, the failure of coordination runs present to future, whereas cases of temptation are failures of coordination that run present to past). The ambiguity in Watzl’s view is that attention functions correctly relative to S, but not relative to P. So, if we assess the performance of attention relative to one’s present priority structure, then there is no failure of attention. Additionally, insisting that we assess attention relative to the initial (commitment-oriented) priority structure misses the point. What explains the performance breakdown here concerns the causes of abandoning the resolution. The abandonment does not result from a failure to integrate or process information; instead, it seems to result from a failure to live up to one’s commitments. Hence, assessing attention relative to the initial priority structure comes at the expense of having the failure of attention be explanatorily relevant to the performance failure.

2.5.3 Broadcast Theories of Attention

Some attention theorists describe attention as a broadcast mechanism that makes information more widely available for consumption in higher-order cognitive processes. This, of course, overlaps with Broadbent’s notion of attention as a bottleneck, but the difference is that broadcast theories do not think attention is needed to manage capacity limitations on information-processing channels.

Smithies (2011) provides a functional characterization of attention as whatever plays the role of making information rationally accessible (see Smithies, 2011: 257). Information is rationally accessible, on his view, when and only when that information
can be used to justify either belief (as in cases of demonstrative thought) or action. This is similar to Prinz’s (2012) view. For Prinz, attention changes information flow and makes information available for further processing. Attention is the mechanism that makes processed contents accessible to working memory (Prinz, 2012: 92). Prinz’s thesis is a bit more metaphysically bold, but his proposal shares a common feature with Smithies, namely that the function of attention is to make information available for wider consumption in an agent’s psychic economy.

While failures of broadcasting are part of a performance breakdown, this, again, does not seem to be a fundamental component of the explanation. On broadcasting views, attention does not determine which bit of information gets broadcast, it is just the mechanism whereby certain bits of information are broadcast. The explanation of a performance breakdown should refer to whatever executive functions are implicated in determining the content of the broadcast, not the mechanism for broadcasting. For this reason, broadcast theories of attention do not supply a notion of failure of attention relevant to explaining performance breakdowns.

2.5.4 Selection for Action Theories of Attention

Wayne Wu (2014) has recently argued for the view that attention is selection for action:

(SfA) S attends to X if and only if S selects X for action (Wu 2014: 96).

For Wu, selection for action is the process of coupling inputs (stimuli) to outputs (behavioral responses) within psychological space (Wu 2014: 80). Psychological space is constituted by psychological relations that obtain between some of an agent’s personal-level mental states and the range of accessible inputs and outputs (Wu 2014: 81).
A core thesis of Wu’s theory of attention is that attention solves problems set by constraints on action. Human beings, for instance, live in a complex world and possess a complex array of skills. Wu, following influential suggestions by Allport (1987) and Neumann (1987), thinks that attention is the mechanism that specifies determinate ways to deploy our skills to achieve a wide range of behavioral outcomes.

Whenever an agent is in a scenario where there is a range of stimuli that are relevant to a range of behavioral outputs, the agent faces a selection problem (Wu 2014: 81). To solve the selection problem, the agent must select an input to guide production of an output. The terminus of this process is action. And Wu argues that attention solves the Selection Problem (i.e., attention is the thing that maps determinate inputs to determinate outputs). This is the basis of Wu’s argument for (S). Whenever an agent selects some item for action, the agent solves the selection problem. And solving the selection problem is the function of attention, so selection of some item involves attending to that item.

Performance breakdowns are difficult to explain on the selection for action theory. On that view, one must explain failure to carry over or failures of foresight by appealing to some failure of appropriate coupling (either failing to couple some input to an output or incorrectly coupling a particular input to a particular output). But that explanatory schema does not seem to map onto any of the examples mentioned in the previous section; rather, the performance breakdowns describe cases where an agent fails to recognize or apprehend the contours of the behavioral space being navigated. It seems more natural to describe these cases as failures of recognition or noticing.

This gets to a problem at the root of (Wu’s version of) the selection for action theory of attention. For Wu, attention emerges to solve the problem of navigating
behavioral space. Psychological relations between an agent, inputs, and outputs constitute and structure this behavioral space. In some cases, an agent is confronted with a behavioral space that contains multiple inputs, multiple outputs, or both. The agent faces a selection problem, namely which inputs to select for which outputs? Attention solves this problem by coupling a particular input to a particular output. Attention, on this view, is needed to navigate behavioral space.

What performance breakdowns illustrate, though, is that agents do not only need to navigate pre-existing behavioral spaces. Sometimes, they face the problem of constructing behavioral spaces. Thus, the problem with the parent who leaves his child in the car is not that he fails to couple inputs and outputs appropriately; rather, he fails to recognize or apprehend a relevant input (his daughter’s being in the backseat), thereby failing to construct an accurate behavioral space. The same applies to double booking. In double booking, one does not fail to select appropriately; rather one fails to see that a relevant input constrains deliberation about current options for coupling. However, similar to discussing other theories of attention, it seems that performance breakdowns will regularly co-occur with failures of attention, but the failure of attention is not explanatorily relevant to the performance breakdown.

2.5.5 Failures of Attention and Performance Breakdowns

This concludes the survey of current theories of attention. The process could continue indefinitely, as there are a wide variety of theories of attention and no shortage of subtle variations within each theory. The current summary, however, gives us a good sense for why failures of attention are, in general, not the right sort of phenomenon to explain the occurrence of performance breakdowns.
The general pattern that recurs in this survey of attention is that something prior to attention seems explanatorily relevant to performance breakdowns. Whether it’s the process implicated in generating cognitive unison, imposing priority structures, or dictating information broadcast, on every theory there is something behind attention that seems to explain performance breakdowns. In general, it seems that whatever is responsible for \textit{shifting} or \textit{deploying} attention is also explanatorily relevant to performance breakdowns.

This also relates to a curiosity about failures of attention. The term ‘failure of attention’ is ambiguous, and that ambiguity seems to generate the sense that failures of attention might relate to performance breakdowns. After all, what is a failure of attention? Individuals are always attending \textit{to something} (assuming they are awake). And a failure to fulfill the function of attention seems akin to a mechanical failure, not something that could ground a performance breakdown. So if the disambiguation of ‘failure of attention’ yields ‘failure of attention to fulfill its function’, then we haven’t generated anything explanatorily useful. Perhaps some people might think that attention is a voluntary activity (as in Watzl, 2017: ch. 7). So perhaps the term ‘failure of attention’ denotes something analogous to whatever the phrase ‘failure to raise one’s arm’ denotes. But this is not correct. \textit{Shifting} attention or \textit{redirecting} one’s attention is sometimes voluntary, but \textit{attending} itself is not a voluntary activity. So we might disambiguate ‘failure of attention’ as ‘failure to shift attention’. But attention shift is not part of the function or activity of attention. Attention shift results either from the influence of bottom-up factors (as in cases of attentional capture) or top-down regulation. These kinds of considerations, coupled with the brief survey of specific theories of attention, suggest
that failures of attention are not the right kind of thing to explain performance breakdowns.

2.6 Vigilance and Performance Breakdowns

Let’s consider, again, the variety of performance breakdowns. Failure of recall and failure to follow through were both instances of failure to carry over, while double booking and coordination failure were instances of failure of foresight. There are interesting parallels between these cases. Failure of recall and double booking are failures of monitoring, whereas failure to follow through and coordination failure are both failures of maintenance. Double booking and coordination failures are *future-directed* failures, while failure of recall and failure of follow through are *present-directed* failures. This reveals two things. First, monitoring and maintenance are both characteristics of the function of vigilance (which, in turn, accurately captures the fact that the performance breakdowns are failures of vigilance). Second, there are temporal components to vigilance. This temporal dimension makes sense. Plans unfold over time and in time. So, if vigilance is meant to be a capacity for planning-relevant information management, then vigilance must have future-directed (‘over time’) and present-directed (‘at a time’) aspects. Thus, vigilance has two dimensions—monitoring and maintenance—that each have temporally variegated aspects.

The maintenance dimension of vigilance overlaps somewhat with self-regulatory constructs like self-control and willpower, so we’ll leave off discussion of maintenance
until the next chapter. Here, I want to focus on the monitoring dimension of the function of vigilance. What can we learn about monitoring from double booking and failure of recall? To bring out the lessons, it is useful to frame the discussion in terms of a case and then work up to a discussion of the double booking and failure of recall case mentioned in this chapter. Consider:

*Milk.* Randy is sitting at his desk at the beginning of the workday. He knows that he is out of milk and wants to buy more. He intends to get the milk at a store that is on the normal route that Randy takes to get home. While driving home, Randy becomes distracted thinking about a paper he is working on. Because of this, he doesn’t notice the store. He drives past the store and arrives home without any milk.

Given the adoption of a goal, an individual acquires a task representation to be implemented in the future (e.g., ‘Get milk’). In Section 2.4.2, I noted that people can do one of two things with these future-directed task representations. One is to store the task representation in working memory. This would, phenomenologically, amount to occurrent thought about the task right up until the moment of action. The other is to store the task representation in long-term memory and associate the representation with a cue. Cues can be either event-based (indexed to situations or landmarks) or time-based and either internal (e.g., ‘I’ll leave when I feel tired’) or external (e.g., ‘I’ll get up when I hear the alarm’).

Monitoring is needed when task representations are associated with a cue and stored in long-term memory. This part of the function of monitoring is cue detection.  

13 Putting the point this way makes it sound like there are clean functional divisions between monitoring and maintenance. This is too simple. The monitoring and maintenance components dynamically interact. But, insofar as we’re rolling out the various components of vigilance, it seems appropriate to consider each component separately before determining the dynamic interactions between these components.
Let’s consider, first, the case of present-directed monitoring. When the associated cue is detected, the task representation is recalled and implemented. Implementation, here, consists in causing regulative mechanisms to configure lower level information-processing pathways in ways that support achieving the task. So, successful cue detection issues in control signal specification and subsequent task performance. In the Randy example, this means that successful monitoring issues in detecting the store, which triggers the representation to get milk. When Randy fails, it might be due to a failure of monitoring (more precisely, when Randy’s mistake is a failure of vigilance, the mistake is due to a failure in monitoring).

Monitoring takes in information about the environment, not just information related to the associated cue. Hence, monitoring might detect changes in the environment that signal shifting task demands (or expected shifts in task demands) that then alter storage of the task representation. If the agent moves into a particularly noisy or demanding environment (or anticipates a movement of that kind), this might trigger a shift of the task representation into working memory, as the agent detects that the task is under threat of being swamped or substituted given the informational demands of the action context.

There is no essential connection between consciousness and monitoring. Indeed, given the limited capacity of consciousness (i.e., the limited number of things about which one can be conscious at any given moment), the function of monitoring is to engage in off-line goal maintenance. If monitoring required consciousness, it would be subject to the same kinds of computational costs and limits as more proactive, working memory-related forms of goal maintenance.
Someone might object that this fails to match evidence about the cost of prospective intentions. Within the prospective memory literature, there is some debate about the cost of prospective memory. Rebekah Smith (2003) found that encoding a prospective intention always incurs a cost on ongoing task performance (measured in reaction time increases). Later work (Smith and Bayen, 2005) explains this cost in terms of prospective intention drawing on preparatory attention mechanisms (i.e., monitoring mechanisms) that are typically deployed solely toward ongoing task performance. This interpretation, however, contradicts the dominant position within the prospective memory literature, namely that prospective memory cue detection is either costly or spontaneous, where spontaneous recall does not incur any costs to ongoing task performance (see McDaniel & Einstein, 2000). Under certain conditions, prospective memory task recall does not draw on preparatory attention mechanisms, but rather relies on bottom-up, associative mechanisms for recall. However, Smith, Hunt, McVay, & McConnell (2007) find that even in the conditions posited for spontaneous recall, there are costs to ongoing performance. Thus, it seems that monitoring (as a function of prospective memory) never fully generates off-line goal maintenance.

The empirical issue here is complicated in two ways. First, Smith and colleagues do not clearly match all of the conditions posited by proponents of spontaneous prospective memory recall (see Einstein & McDaniel, 2007 for comment). Second, there are conceptual reasons to think that the conditions for spontaneous recall have not been well understood. Briefly, according to a shared resources model of cognitive control (see Allport, 1989; Cohen, 2017), capacity limits on control derive from the fact that the brain utilizes a limited number of general task representations that are shared among a variety
of different cognitive processes. This view predicts that prospective memory tasks incur a cost to ongoing task performance in virtue of utilizing a representation shared with processes implicated in performing the ongoing task. When the prospective memory cue is not shared with the ongoing task, there should be no cost. The lesson to draw from this, then, is that prospective memory recall is likely spontaneous under certain conditions, showing that individuals can engage in off-line goal maintenance. Because the difference between costly and spontaneous recall is still poorly understood, it is difficult to tell the extent to which ordinary prospective memory is costly.

In present-directed cases of the monitoring dimension of vigilance, the agent monitors for cues associated with task representations stored in long-term memory and monitors for changing environmental conditions that require a shift in control strategy. Something similar happens in future-directed cases of monitoring. Plans are supposed to constrain future deliberation. When an agent engages in future-directed monitoring, the aim is to bring to mind those plans in a way that constructs the deliberative space accurately. This does not occur in double booking, where the agent fails to bring to mind some consideration that ought to constrain her deliberation. The same distinctions that characterize present-directed monitoring also hold with future-directed monitoring.

When we reflect on these performance breakdowns, we get a clearer sense of the functional profile of vigilance. As we’ve seen, successful monitoring is a function from plans to accurate apprehension of the action context (whether that action is bodily movement or deliberation). This accurate apprehension consists in detecting cues associated with goals stored in long-term memory and information updating (typically related to changes in the environment that demand shifts in strategy). In other words,
successful monitoring consists in accurately constructing one’s action context and
successfully shifting strategy when needed.

This might lead to some confusion. What, after all, is a failure of vigilance? And, won’t these failures of vigilance just be glitches? In what sense is the agent doing anything wrong? We need an account of the maintenance dimension of vigilance before giving a satisfying answer to these questions. But, as a first pass, we can say the following. Failures of vigilance along the monitoring dimension occur in one of two ways. Either there is improper control strategy selection at the task representation storage stage, or there is goal neglect (either from goal substitution or mind wandering). Thus, we might think that a monitoring failure exhibits a lack of focus. Increasing one’s focus, however, consists in selecting an alternative control strategy. Goal neglect, on the other hand, is more complicated. Consider the milk case again. Roughly, Randy might construe his goal in a way that leaves it susceptible to substitution from entrenched defaults (habits). That is, he might construe his goal as ‘Get milk and get home’ (‘A & B’). But this is easy to break into two goals: ‘A’ and ‘B’. Since ‘A’, here, represents a task that deviates from one’s routine (namely, the tasks represented as ‘B’), this way of construing the goal leaves Randy susceptible to substitution. This might seem like a failure of maintenance, but that’s not necessarily the case. As we’ll see in the next chapter, Randy might exhibit a failure to monitor in that he overestimates the probability that he will successfully achieve the task of getting milk given a certain control strategy selection. And that overestimation could be characterized as a failure of monitoring. Again, we need the account of maintenance (and the computational principles that govern exercises of vigilance) on the table before we can fill out the details (which we’ll get in Chapter 3).
However, we can get a sense for what Randy’s failure of vigilance consists in such that the failure is not a mere glitch that occurs at the subpersonal level.

But let’s leave aside foreshadowing and get back to the main thread. The account offered here shows that a successful exercise of vigilance in the monitoring dimension consists in an agent accurately construing her action context in light of previously formed plans. In the present-directed aspect, this takes the form of apprehending considerations relevant in light of one’s prior plans. In the future-directed aspect, this takes the form of apprehending constraints on deliberation relevant in light of one’s prior plans. The unifying element, in both cases, is that the agent moves from plans to context. A failure of vigilance consists either in an agent failing to select the appropriate control strategy for managing prospective intentions or in an agent overestimating the probability that she would follow through on her plans.

One might object that this account reduces monitoring to a purely instrumental process. Don’t we sometimes monitor for better goals, and not just considerations relevant to our already adopted goals? This objection feels like a demand for an account of virtuous vigilance, which lies outside the scope of the present essay. It is correct to say that one’s goals and plans determine the shape of properly functioning vigilance, and properly functioning vigilance is morally neutral (i.e., properly functioning vigilance can serve good or evil plans). However, we can do a little more than just pass the buck here. Briefly, agential attitudes seem to divide naturally into commitments, goals, and tasks. Tasks serve goals, and goals are means of realizing a commitment. As I see it, commitments are quite coarse-grained (‘be a good husband’ or ‘be environmentally conscientious’). There are multiple ways to realize a commitment, and the goals we adopt
are specific ways of realizing those commitments (‘buy flowers twice a week’ or
‘volunteer to clean the local river’). Tasks are then means of achieving the goal. The
discussion so far has focused on monitoring relative to goals. But we could monitor for
goals relative to our commitments. This dimension of monitoring would map most
naturally to the role that monitoring plays in deliberation, insofar as monitoring relative
to goals makes most sense when one is deliberating about changing goals. I do not think
there can be monitoring for better commitments. The issue here is complicated, but I
adopt a roughly Frankfurian line that commitments are constituents of the will and form
the background against which we make any choices (see Frankfurt, 1993). Hence, there is
no deliberation about commitments and, consequently, no way to change our
commitments as a result of exercising control.¹⁴ This implies that vigilance plays no role
in selecting or shifting commitments. The supporting role characteristic of vigilance
mentioned in Chapter 1 partially explains why vigilance is inert relative to commitments.

This characterization of vigilance also connects the monitoring dimension of
vigilance to theories of attention. As we saw above, vigilance and attention are distinct.
However, a theory of vigilance ought to explain two things: (1) why people often ascribe
functional properties of vigilance to attention, and; (2) the relationship between vigilance
and attention in a typical-functioning adult human being.

The problems with discussing attention recur at this point. The discussion here
will be brief, and should be taken as suggestive of lines for future research. Two factors, I

¹⁴This might seem overly strong. What about cases where someone discovers conflicts between
commitments? There are two responses to this. First, commitments are vague and indeterminate. Thus, it is
hard to imagine realistic cases where commitments themselves would come into conflict (there might be
cases where particular goals adopted relative to different commitments come into conflict, but that’s a
different case). Second, the discovery of conflicts among commitments might cause one to temporarily
suspend one or either commitment in subsequent deliberation about how to revise one’s commitments.
think, contribute to people confusing vigilance and attention. The first is that attention is a paradigmatically present-directed process or cognitive capacity. Attention does not unfold over time and can operate independently of plans, goals, and other temporally extended aspects of our agency. The second factor is that the fact of our planning agency, and the subsequent need for temporally extended cognitive capacities, results in people ascribing to attention properties that make it suitable for functioning within a planning agent. As we have seen, there are good reasons to think that vigilance and attention are distinct, and that vigilance regulates attention in accordance with the demands of planning agency.

There are, however, natural fits between this account of vigilance and other accounts of attention. In the discussion of theories of attention, from above, we saw that attention stands in need of a regulatory mechanism. Thus, on broadcasting views, there needs to be some mechanism that determines what information to broadcast for further consumption. Vigilance seems to play this role, insofar as it determines that plan-relevant information be broadcast or shifted into/out of working memory. On a unison view, there needs to be some process or mechanism that brings one’s cognitive resources into unison. Since vigilance is a regulatory mechanism implicated in coordinating information extended over time, it makes sense that vigilance would play some role in achieving cognitive unison. Finally, on a selection view, attention functions as a mechanism that binds inputs to outputs. At a level of abstraction, attention functions to enable an agent to move from apprehension/appraisal of an action context to action. This fits naturally with the characterization of vigilance mentioned above, as the functional characterization revealed that vigilance (at least in its monitoring dimension) functions to enable an agent
to move from plans to accurate apprehension/construction of an action context. As noted above, one problem with Wu’s selection theory is that agents sometimes need to construct the behavioral spaces through which attention enables them to move. In short, the function of vigilance is to construct behavioral space, while the function of attention is to navigate that behavioral space.

The shortcoming of the present survey is that theories of attention vary widely, as already noted. However, the brief discussion here should show certain patterns about how vigilance connects to attention. In particular, vigilance regulates attention, which, in turn, explains why attention manifests vigilance.

2.7 Empirical Studies of Vigilance

Thus far, I’ve argued that vigilance is not a capacity for sustained attention or any kind of attention. Instead, attention always manifests vigilance. This proposal might appear to conflict with the vast majority of empirical studies of vigilance because key results in vigilance research are explained by appealing to features of attention. To see this, we need a brief look at the history of empirical research on vigilance.

Empirical studies of vigilance began in earnest in the 1940s when N.H. Mackworth developed the Clock Task to study vigilance (Mackworth, 1948; there were earlier studies of vigilance, dating back to Head (1923), but these earlier studies were infrequently referenced after Mackworth’s research). The Clock Task was modeled after the activities of Naval radar operators during WWII, but in reality the Clock Task was much simpler than its real-world counterpart. Participants faced a large clock with a moving hand. The standard Mackworth clock has a 15cm pointer that moves in 100
discrete steps of $3.6^\circ$ around the face, with each step occurring within a 1-sec. interval. Occasionally, the clock would double jump. Participants were given simple instructions: immediately report double jumps by pressing a button. In Mackworth’s original paradigm, the double jumps occurred randomly over long intervals (~2-3 mins.) for very long trials (1.5-2 hours; this was the shift length for a typical radar operator). Mackworth discovered that after about 15 minutes, participants started missing double jumps. After this point, they would miss incrementally more double jumps over time, with performance degrading smoothly over the course of the trial. This performance deficit came to be known as the vigilance decrement, and the empirical study of vigilance focused on identifying the factors that affected the onset and progression of the decrement.

Over time, researchers labeled the Clock Task a test of sustained attention and described vigilance as a capacity for sustained attention. This provides an easy explanation of the decrement. The capacity for attention is limited and consumed on tasks like the Clock Task. As attentional resources are depleted, focusing attention on a task becomes more difficult. When attentional resources are fully depleted, performance drops off. Versions of this ‘resource depletion’ model of vigilance have been offered by a number of different researchers (Grier et al., 2003; Helton et al., 2005; Helton & Warm, 2008; Helton & Russell, 2011; See, Howe, Warm, & Dember, 1995; Temple et al., 2000). This explanation applies not only to the Clock Task, but to any continuous stimulus discrimination task (this includes any uncued Go/No-Go paradigm, which is the most common task paradigm for studying vigilance; see Davies & Parasuraman, 1982).
Several findings, however, do not fit well with the resource depletion model of the decrement. For one, having the relevant stimulus (e.g., double jumps) occur at regular, predictable intervals nearly eliminates the decrement (Robbins, Milstein, and Dalley, 2004). Introducing warning signals that occur just prior to double jumps nearly eliminates the decrement (Maclean et al., 2009). And, making the stimuli more ecologically valid (or subjectively interesting) nearly eliminates the decrement (Barron et al., 2011). This suggests that people are not deploying sustained attention on task to the point of depletion; instead, it suggests strategic engagement with a task that takes time to complete.

Thus, one way to reinterpret empirical studies of vigilance is to recognize that participants strategically engage with inherently boring tasks. They engage, disengage, and then re-engage. Periods of disengagement increase the likelihood of missing double jumps, which suggests an explanation of the decrement. People deploy attention to the task long enough to form and internalize a task schema (i.e., a set of rules that map inputs to outputs). With the schema formulated, there is no further benefit of maintaining attention to the task, so participants disengage.

On this view, what explains the decrement is a strategic allocation of cognitive resources. Vigilance tasks, like the Clock Task, do not require much attention or effort to successfully complete. So people shift their attention elsewhere. This shows that the vigilance decrement is not a product of depleted resources. It also shows that something shifts attention based on implementing a strategy for managing cognitive resources. And this, recall, is the function of vigilance in planning agents. Vigilance tasks, then, measure
vigilance by putting people in situations that require strategically managing one’s

cognitive resources.

This rough sketch outlines how this theory of vigilance might be applied to the

empirical study of vigilance. What’s left out is a story about how vigilance interacts with

motivation and memory. That, however, requires thinking about the costs of control,

which we’ll explore in the next chapter. The full empirical framework of vigilance is left

for Chapter 6.
CHAPTER 3:

VIGILANCE AND SELF-CONTROL

3.1 Introduction

The standard story about self-control centers around three claims. First, many think that self-control is limited. Exercising self-control at one moment makes it more difficult to exercise self-control thereafter in the absence of a recovery period. There are various explanations for these limitations, ranging from consumption of limited resources (Baumeister et al., 1998) to structural and functional limitations of processes supporting self-control (Edin et al., 2009; Ma & Huang, 2009; Miller, 1956; Oberauer & Kliegel, 2006; Usher, Cohen, Haarmann, & Horn, 2001). Second, exercising self-control requires effort and is, for that reason, aversive (see Kool et al., 2010; Kurzban, 2016; Inzlicht, Shenhav, & Olivola, 2018). Finally, the function of self-control is to resist temptation or suppress impulses (Ainslie & Haslam, 1992; Bargh & Chartrand, 1999; Holton, 2009; Joosten et al., 2015; Mischel et al., 1989; Muraven & Baumeister, 2000; Myrseth & Fishbach, 2009; Soutschek et al., 2016; Watson, 1977).

Discussions of self-control are historically tied to explanations of weakness of will (Mele, 2011). Framed in this way, the standard story of self-control makes sense. In such a case, the agent must employ self-control to avoid succumbing to temptation (functional claim). Because of the temptation, the agent must use self-control to resist
something she currently desires (aversion claim). Lastly, if the individual does not remove herself from the tempting scenario, then she eventually gives in because her self-control is depleted (limitation claim). It’s a nice package. And it’s incorrect.

We can challenge the standard story from the bottom up. While self-control is sometimes employed to resist temptation or suppress impulses, this certainly does not exhaust the range of applications for self-control. To see this, consider the following example:

*Hot Car.* Kate normally dropped her daughter Sam off at daycare before working from home. But one summer morning, Kate had an early meeting and her daughter was sleeping in. This meant Brad—Kate’s husband—had to take Sam in before heading to work. Unfortunately, Brad forgot to drop his daughter Sam off. By the time he remembered, the little girl had been sitting inside a car, windows closed, on a summer day in Texas for several hours. EMT’s declared Sam dead a mere 80 minutes after Brad called 9-1-1.

This appears to be a failure of self-control. The case describes an agent (Brad) that forms a plan to accomplish several goals. Brad intends to carry out each element of the plan. In the end, Brad fails to follow through on executing each part of the plan, thereby failing to achieve all of the goals he aimed to achieve. Further, it appears that the process that disrupts Brad’s following through on his plan is something internal to him. There are none of the characteristic mad scientists or evil demons of philosophical fiction afflicting Brad. Brad himself just failed to focus. These considerations support describing the case as a failure of self-control. But where’s the temptation? Where’s the impulse? You won’t find them because they aren’t there. This suggests that there is more to self-control then the motivational dimension encapsulated in the functional claim of the standard story.
Self-control is not exclusively aversive. Theorists typically infer aversion from the fact that exercising self-control requires effort. But neither self-control nor effort are necessarily aversive. Consider the fact that learning a new language or a new instrument requires lots of self-control and effort, but neither experience is aversive (or, at least, not always aversive). The aversion claim ignores a general truth, namely that some of the enjoyable experiences in life also require lots of effort.\(^\text{15}\)

Discussions of limitations on self-control are complicated, and we should distinguish the target of the discussion. There is the phenomenon of self-control limits and the candidate explanations of those limits. Let’s consider first the phenomenon of self-control limitations. In a standard study on self-control limits, participants in the experimental condition perform two tasks, both of which require self-control. The control group performs a non-control task in the first block and a self-control task in the second block. These tasks range from Stroop tasks to emotion suppression tasks. Participants in the experimental condition typically perform worse on the second task (or tasks subsequent to an initial self-control task) across a variety of dimensions relative to the control group (Schmeichel, 2007; see Hagger et al., 2010 for a meta-analysis of depletion effects in self-control studies). Some of the findings include: lower inhibition after performing a controlled task (Vohs & Faber, 2007), stronger bias toward behaviors that require only automatic processing (Schmeichel, Vohs, & Baumeister, 2003), higher rates and intensity of prejudicial judgments (Muraven, 2008), and diminished performance on tasks that require a speed/accuracy tradeoff (DeWall, Baumeister, Mead, & Vohs, 2011).

\(^{15}\) Spinoza captures this point eloquently: “How would it be possible, if salvation were ready to our hand, and could without great labour be found, that it should be by almost all men neglected? But all things excellent are as difficult as they are rare” (Ethics Vp52 s).
These depletion effects have shown up in hundreds of independent studies (see Baumeister & Vohs, 2016: 75-93 for a summary of findings). A meta-analysis showed that effects from these various studies were both robust and significant (Hagger et al., 2010). However, some challenge whether this meta-analysis appropriately accounted for possible publication and small-study biases (Carter & McCullough, 2014). Following up on this possibility, a later meta-analysis found a corrected effect that is effectively zero (Carter et al., 2015). And, though a subsequent preregistered, multi-lab replication effort found significant effects (Sripada et al., 2016), a meta-analysis of the results found an ego-depletion effect close to zero (Hagger et al., 2016), and Bayesian analysis of the results supported this null finding (Etherton et al., 2018). Additionally, attempts to replicate some of the training effects on ego depletion found in the original Hagger et al. (2010) meta-analysis failed when correcting for time-scale and task type (Miles et al., 2016). When participants performed controlled tasks that lasted over longer time-scales (i.e., tasks with higher ecological validity), Wenzel et al. (Manuscript) found a reversed ego depletion effect.

What this means is that ego depletion effects are more complicated than the original findings from two-task experimental conditions suggested. While there is some evidence that ego depletion effects do not exist, it seems more probable that ego depletion effects are highly sensitive to contextual factors. This would explain the highly variable findings. In fact, in what is becoming a near weekly tradition, yet another meta-analysis of ego depletion studies found significant overall effects of depletion across nine different task types (Dang, 2018). Hence, it seems safe to say that ego depletion effects are real, so we need some explanation for them.
One explanation is that exercising self-control consumes a limited resource. Hence, an exercise of self-control depletes the agent’s resources. When this occurs, a central governing mechanism registers that the agent is consuming resources at an unsustainable rate, thus causing the brain to reduce exertion and conserve energy (see Evans, Boggero, & Segerstrom, 2016). This framework construes self-control as analogous to a muscle and willpower as analogous to physical strength and posits that the limited resource is metabolic (see Baumeister & Vohs, 2016).

There are two problems with this view. The first is that the Strength Theory has difficulty explaining the variety of contextual factors that modulate ego depletion effects (see Kurzban et al., 2013). The second is that there appears to be no plausible metabolic substrate that supports self-control. The original proposal for the metabolic substrate was glucose (Gailliot & Baumeister, 2007; Gailliot et al., 2007). This proposal has not fared well over time. Later studies found that exerting self-control sometimes raised blood glucose levels over time (Ainsworth, Baumeister, & Boroshuk, Manuscript). A meta-analysis of various glucose-related results found no significant effects (Dang, 2016). The proposal is also highly counterintuitive. Brain processing consumes a miniscule amount of glucose relative to the rest of the body. Exercising self-control raises glucose consumption by a miniscule fraction. If these small glucose consumption increases are sufficient to trigger depletion effects, then self-control would be much too volatile to support temporally extended agency (a point also raised by Wenzel et al., Manuscript).16

---

16 Feng et al. (2014) point out that self-control processes depend on prefrontal cortex. However, other cognitive and perceptual processes (like vision) engage the prefrontal cortex for sustained periods of time without any depletion effects on the scale observed in self-control studies.
In the wake of suspicion surrounding the Strength Theory, a number of alternative theories about the nature of self-control have emerged. Many of these alternatives cite motivation and attention as essential components of self-control (e.g., Botvinick & Braver, 2015; Eastwood, Frischen, Fenske, and Smilek, 2012; Inzlicht, Schmeichel, & Macrae, 2014; Hockey, 2011; Inzlicht, Shenhav, Olivola, 2018; Kool & Botvinick, 2014; Kurzban et al., 2013; Shenhav et al., 2017). There are some important differences between these theories, but the central thread connecting them is that ego depletion effects result from shifts in motivation (thereby causing shifts in attention). For example, Inzlicht, Schmeichel, and Macrae (2014) explain that one explanation of ego depletion effects in laboratory studies is that participants are motivated to think about other things, like what they’ll be eating for dinner later or what they’ll do with their friends over the weekend. This also explains why shifting incentives mitigates the ego depletion effect.

Limitations of self-control, then, reflect our fleeting motivations to carry out any single task (and, perhaps, our underlying propensity to shift attention from tasks we have to do to tasks we want to do).

While the Motivational Theory avoids the problems that plague the Strength Theory, the view is also unintuitive in some respects. Consider, again, the *Hot Car* case. There is good reason to think that the case describes a failure of self-control, but there’s no identifiable shift in motivation that explains the self-control failure. In fact, we can press this point and say that it’s *wildly implausible* that parents in these kinds of scenarios have changed their preferences or shifted motivation. That would imply that the parent has come to value getting to work on time over the life of their child. We see this inadequacy when we extrapolate out to other performance breakdowns. For example,
forgetful cooks do not suddenly lose motivation to have their kitchen not be engulfed in flames. And forgetting a birthday or anniversary does not reflect an individual’s devaluation of a relationship.

Both of the major theoretical approaches to self-control turn out to be inadequate. Thus, we need some approach that explains apparent limitations on self-control (and suggests some principled explanation of the influence of contextual factors on these limitations) in an empirically tractable way that maps onto the various dimensions of self-control. Additionally, it would be nice for the theory to explain the connection between effort and self-control in a way that avoids construing self-control as necessarily aversive and to explain where temptation resistance/impulse control fit into the scheme of self-control. Finally, we want our resulting theory of self-control to explain what’s going wrong in the Hot Car case.

In this chapter, I propose a self-control as vigilance theory of self-control. In Section 3.2, I explain why we should connect self-control and vigilance. In Section 3.3, I propose an alternative way to understand limitations on control that grounds self-control limits in efficient cognitive network architecture. Section 3.4 discusses the computational principles that govern control allocation. With this in hand, in Section 3.5 I outline the maintenance dimension of vigilance and connect this to the computational account of control allocation from Section 3.4. Finally, Section 3.6 ties together various threads, discussing the connection between self-control and effort, the kind of self-control that impulse control/temptation resistance requires, and the explanation of various ego depletion effects.
3.2 Self-Control as Vigilance (I)

To see the reasons for connecting self-control to vigilance, let’s step back and consider what the purpose of self-control is. If we reject the claim that the function of self-control is just to resist temptation or impulse, then why do creatures like us have self-control at all?

One of the persistent underlying themes of this project is that human beings pursue a variety of goals and commit to a wide range of projects. Living a meaningful life requires heterogeneity of pursuits. However, these goals cannot be simultaneously realized or even simultaneously pursued. Thus, multiplicity of goals, combined with limits on cognitive and bodily resources, generates the need for a goal maintenance mechanism, i.e. vigilance. Vigilance regulates goal-relevant information and the implementation of goal-relevant information in mind that facilitates goal-directed action.

Notice that a perfectly virtuous creature with heterogeneous pursuits and certain cognitive limitations that experiences no temptations and is subject to no impulses would still require vigilance for effective goal maintenance. Further, it seems intuitive to think of part of the function of self-control as goal maintenance. Self-control, then, amounts to the pursuit of multiple goals in an effective and efficient manner.

One benefit of this account is that it classifies obsessions as failures of self-control. This is not possible on the standard story of self-control. The obsessed individual doesn’t succumb to temptation or impulse; rather, she focuses too much on one thing at the expense of everything else. Obsession, then, manifests lack of efficiency and, hence, lack of self-control to some degree.
Given the overlap between the functional description of vigilance and the functional description of self-control, it makes intuitive sense to consider the claim that self-control is a species of vigilance.\textsuperscript{17}

3.3 The Limits of Control

If we want to construct a theory of self-control that departs from the standard story, I suggest we start by reexamining the limitations on self-control. We can start with a simple question: Why is it that an Apple Watch—with four circuits—can calculate 2 two-digit math problems simultaneously, but the human brain—with over 20 billion cortical neurons—cannot calculate 2 two-digit math problems simultaneously? As the question suggests, control limitations derive neither from metabolic limitations (extra sugar won’t help) nor from structural/functional limitations (as a structurally and functionally limited computational device can do things the human brain cannot).

The Stroop task provides an instructive example about the limits of control. Try as hard as you might, you cannot engage in color identification and word naming simultaneously (if you could, incongruent Stroop tasks wouldn’t be harder and take longer than their congruent counterparts). You cannot imagine what it’s like to be in Berlin and what it’s like to be in Bogotá simultaneously. The examples are endless. But, again, it does not seem like these limitations are ultimately reducible to structural or functional limitations. The human brain has an enormous amount of resources that it could devote to controlled processing. So, even if structural or functional limitations are

\textsuperscript{17} Self-control is just a species of vigilance, not equivalent to it. This is because some exercises of vigilance might not consist in a manifestation of self-control. Again, this will be a point of discussion in Section 3.4.
part of the explanation for limited controlled processing, it seems that something else is fundamentally explanatory.

But, if we’re not going to appeal to metabolic, structural, or functional limitations, then what might explain these strict limitations on control? To get to the answer, consider an example. We utilize task representations to guide goal-directed behavior. So, we store task representations that correspond to crossing the street, whisking the eggs, and watering the garden. However, you wouldn’t want a unique representation for every unique task you can perform. As mentioned in Chapter 2, you don’t need a separate street-crossing task representation for every street you happen to cross. A single, generalized ‘street-crossing’ task representation will do (see Rougier et al., 2005).

This is true in general. A basic stock of representations that can be utilized for multiple tasks is more efficient than a highly specialized basic stock of representations. Utilizing a basic stock of generalized representations that can be flexibly deployed across a variety of tasks is known as multiplexing. Allport, Antonis, & Reynolds (1972) were the first to suggest that the brain multiplexes. Multiplexing, however, introduces the possibility of channel cross-talk. If you have a variety of available input-output task mappings subserved by the same representation, then these tasks might potentially interfere with each other (Forbus, Gettner, & Law, 1995; Hinton, McClelland, & Rumelhart, 1986). The Stroop task provides a simple example. The input-output mappings that correspond to word naming and color identification both utilize the same representation. Hence, the processes that subserve these two tasks cannot be activated simultaneously and activating one implies performance deficits for the other. This reveals that multiplexing correlates negatively with the capacity to multitask, as the more
multiplexing occurs, the greater the possibility of cross-talk, which reduces the capacity to multitask without possibility of interference (see Feng et al., 2014 for a computational model of the absolute limits of multiplexing on multitasking in an optimal control network independent of network size and number of control nodes).

Dramatic limitations on control reflect the brain’s preference for efficient coding through multiplexing. We can infer the brain’s preference for multiplexing over multitasking from the fact that we can perform so few controlled tasks simultaneously. This, however, generates a problem of interference. A high degree of pathway overlap implies a number of potential sites of interference between task mappings. To solve this problem of interference, you need a control manager. Thus, Jon Cohen claims that: “These [shared resource] models suggest that constraints on the simultaneous execution of multiple tasks can be viewed as the purpose of control, rather than a limitation in its ability” (2017: 5, emphasis original).

The function of the manager unit, as Cohen suggests, is to adjudicate conflicts between overlapping pathways and bias lower-level information processing in ways that support goal-directed behavior (see also Botvinick et al., 2001). Biasing is accomplished by allocating control to unique pathways to alter the threshold for a neuronal population to fire. Hence, conflict management requires allocation of control (we’ll consider the computational principles of control allocation in the next section).

One natural question is why the brain (or the evolutionary principles that govern brain development) did not generate a unique control unit that manages a single point of potential cross-talk. However, the addition of extra control units (beyond the optimal convergence point) actually generates performance deficits (see Feng et al., 2014: Fig.
S5). Both error rate and expected reward attainment decrease in proportion to additional control units. Feng and colleagues do not offer a computational story for the performance deficits, but the effect makes intuitive sense. Too many managers decreases overall efficiency. There might also be adaptive value in having fewer control units. With fewer control units, local and global reconfigurations of task mappings in response to task acquisition can occur more easily. This is because a system that utilizes minimal control units requires a simpler network architecture. Over time, this might represent a metabolic advantage over massively modular architectures (cf. Anderson, 2014: 38).

This framework provides a new way to think about limitations on self-control. Multiplexing, combined with a small number of control nodes, constrains the capacity to multitask. This computational account of control limitations is more empirically tractable than the metabolic account, and also explains the structural and functional limits on controlled processing. Return to the Apple Watch. The watch doesn’t multiplex, instead utilizing discrete informational units for each calculation. This is fine for the watch because it doesn’t do much and so can afford to prefer multitasking. We, however, have the advantage of being able to do innumerable things. The price for this is that we multiplex for the sake of efficiency. This generates pathway overlap that decreases capacity to multitask. However, we can solve this interference problem with control nodes that manage channel interference. This generates a new question: what are the computational principles that dictate allocation of control across these various channels?
3.4 The Expected Value of Control

One cornerstone of cognitive architecture research is that cognitive efficiency is built on a speed/accuracy tradeoff (see Bogacz et al., 2006). There are benefits associated with a system of heuristics and defaults (e.g., fast processing speed), but there are also costs (e.g., error and inflexibility) (Miller & Cohen, 2001). There are benefits associated with a deliberative system that assesses available actions relative to an internal causal model of the environment (e.g., accuracy and flexibility), and there are costs (e.g., slow processing speed).

In light of this, the computational principles that govern control allocation should solve for an optimal balance between speed and accuracy. Additionally, a background presumption here is that the network is configured for long-term maximization of reward. So the optimal balance between speed and accuracy reflects an aim toward maximizing reward. But, given the discussion from Section 3.3, the optimal balance must be measured relative to the bounded computational powers of a control network that utilizes shared representations. One intuitive engineering principle is that the network should rely on a system of default settings, with control units intervening only when necessary. This is plausible because optimal network design would suggest that you use the fast/computationally cheap system as much as possible and correct for errors (with the slower/computationally expensive system) when necessary. Control units, then, would monitor for conflicts among lower-level information processing units. When conflict is detected, the control unit would calculate the costs and benefits of engaging control to resolve conflict.
Recent computational models of control allocation provide a framework for thinking about these issues at the psychological level. That is, if we understand the computational principles that govern control allocation at the neurobiological level, we will gain a foothold for understanding how those principles manifest at the psychological level. One popular model is the Expected Value of Control (EVC) model articulated by Jonathan Cohen (see Shenhav, Botvinick, & Cohen, 2013 for an early statement of the computational and mechanistic aspects of the view). The mechanistic aspects of the EVC are not important here, so I will focus only on the computational elements. The following three computations represent the core of this evaluation (these equations are taken directly from Shenhav, Botvinick, and Cohen, 2013: 221):

**Equation 1:**

\[
\text{EVC}(\text{signal, state}) = \left[ \sum_i \Pr(\text{outcome}_i|\text{signal, state}) \cdot \text{Value(\text{outcome}_i}) \right] - \text{Cost(\text{signal})}
\]

**Equation 2:**

\[
\text{Value(\text{outcome})} = \text{ImmediateReward(\text{outcome})} + \gamma \max_i [\text{EVC}(\text{signal}_i, \text{outcome})]
\]

**Equation 3:**

\[
\text{signal}^* \leftarrow \max_i [\text{EVC}(\text{signal}_i, \text{state})]
\]

The computational principle of the EVC calculates the value of a particular control signal based on three components. The first component represents the probability of achieving a particular outcome given that a control signal is sent in a particular
environment \((Pr(\text{outcome}|\text{signal, state}))\). The second component is the value of the outcome \((\text{Value(\text{outcome},)})\), and the third component represents the intrinsic cost of the signal \((\text{Cost(\text{signal})})\).

Currently, there are no widely accepted theories of which mechanisms compute values for the probability of achieving an outcome given transmission of a control signal nor what computational principles govern these mechanisms. Some suggest that certain mechanisms simulate controlled behavior to rapidly generate estimates (Pezzulo, Rigoli, & Chersi, 2013). Others suggest that model-free learning mechanisms update estimates of probabilities for the system without the use of simulation (Gershman, Horvitz, & Tenenbaum, 2015; Braem, 2017). Others find evidence for the use of heuristics to estimate probabilities without assessing task-specific demands (Dunn, Lutes, & Risko, 2016). At this point, not enough is known about task representation acquisition and updating to know how this process is carried out. But the point remains that the computations underlying control allocation are sensitive to the probability of achieving the outcome. When the probability of success diminishes, control decreases (see Kool et al., 2017).

The representation of value has two parts. The first concerns the immediate expected reward upon achieving the outcome. In itself, this is intuitive. Part of the cost-benefit analysis associated with determining control allocation should be sensitive to the potential benefits of achieving the outcome. However, the EVC model also adds a temporal component to value representation. Outcome value representation reflects the kinds of action-reward sets that will be available from the achieved outcome state \((γ\max[EVC(\text{signal}, \text{outcome})])\). Thus, an immediate high reward outcome will diminish
in overall value if there are no desirable action-reward sets available from the state in which one has attained the outcome.

This aspect of the value representation reflects part of the cost of control (but not the intrinsic cost represented in the third component). Consider two possible outcomes, $O_1$ and $O_2$. Suppose that the immediate reward associated with these outcomes is such that $O_1 > O_2$. But, achieving $O_1$ would require using most of your ‘cognitive juice’, leaving you in a position where you could not pursue other rewards without some ‘rest’ (the terms in quotes might incline one toward an interpretation of limited resource consumption, but that would be incorrect; the issue of limitations is discussed below). Achieving $O_2$, on the other hand, consumes less ‘juice’, and so leaves you in a better overall position to achieve future rewards. Of course, the value of $O_1$ might greatly exceed $O_2$, thereby nullifying these costs. The point is simply that the computation is sensitive not just to the value of available outcomes, but about future deployments of control and the future action context.

Finally, there is the cost of the signal. Recently, there has been debate about how to interpret the cost of control. Roughly, this represents the intrinsic costs of control. Recall that our cognitive network architecture is designed so that only one (or a very small number of) controlled task(s) can be pursued at any one time. When we utilize control, we intervene on the ‘default’ settings of lower-level information processing pathways, thereby configuring the system to support goal-directed behavior in the circumstances. However, this configuration means that any number of available configurations are not available, meaning that the rewards associated with pursuing other
actions are missed. Hence, we can understand the intrinsic costs of control in terms of the opportunity costs associated with not pursuing other goals.

It is worth noting that there are two separate cost values within the allocation computation. The first is part of the representation of the perceived value of achieving the outcome. This cost reflects ‘depleted’ resources that result from using control. This depletion, however, is metaphoric. The real cost consists in task-switching costs associated with reconfiguring the system either toward another goal or back to a default (unconfigured) state. This also implies diminished behavioral flexibility relative to non-goal-related activities, thereby increasing switching costs. The intrinsic cost, on the other hand, reflects the opportunity cost of thinking and pursuing one goal at the exclusion of others or at the exclusion of plural (unconfigured) goal pursuit (i.e., mind wandering).

Hence, the EVC model assigns a unique value to a particular control signal as a function of the probability of achieving an outcome given the transmission of a control signal, the value of the outcome, and the cost of that signal. To see how the model works, consider an idealized, non-iterated incongruent Stroop task condition. Here, one is presented with a word (e.g., ‘green’) filled in with a different color (e.g., RED) and told to identify the color, not the word. When the word is presented, there is a conflict among lower-level information processing units associated with word naming and color identification (since both utilize the same representation). The EVC model describes the calculations that control units compute to determine whether to use control. Because the scenario is idealized, we can assume that the probability of achieving the outcome given the signal = 1, while the probability of achieving the outcome without the signal = 0. The value of the outcome-given-signal is high (as it conforms to experimenter instructions),
so we can assign it a value of 1, whereas the value of outcome-without-signal is 0. The cost of the signal drops out here, as the Stroop task in this case is not repeated. Hence, relative to the experimental condition, there is no need to calculate values relative to the state where one has achieved the outcome. The expected value of the control signal outweighs both its costs and the alternative based on not signaling. With the computation complete, the control unit signals to bias the network toward goal-relevant processing. Of course, the computations get more complicated as probabilities are added, tasks become temporally extended, and the value of other control signals must be weighed. But the idealized example shows how the various components operate.

The EVC model is one of many available models of the cost-benefit analysis associated with control allocation. So why go on at length about the components and implications of this one model? There are good reasons to prefer this model to others. For one, the EVC model has plausible neurobiological realizers and provides an integrative framework for thinking about various neurobiological mechanisms (Braver et al., 2014; Zhang, Stock, and Beste, 2016; Ebitz & Hayden, 2016). In particular, recent work seems to show that we can map EVC computations to various functionally localized neural mechanisms (see Shenhav, Botvinick, & Cohen, 2013 for a review). Also, experimental work supports the EVC model, especially in reinforcement learning paradigms, where the EVC makes divergent predictions from other control allocation algorithms (e.g., Kool et al., 2017). Finally, the model connects with recent theoretical advances in artificial intelligence research and machine learning applications. For example, the EVC model mirrors theoretical frameworks that model rational metareasoning and algorithm selection in artificial intelligence systems (Lieder et al., 2018) and implementing the EVC model.
has produced advances in machine learning (LeCun, Bengio, & Hinton, 2015). In addition, the EVC model seems compatible with other computational models of control allocation (see Kool, Shenhav, and Botvinick, 2017), but that argument lies outside the scope of the present paper. Should the EVC model turn out incorrect, it is likely that whatever the correct model is will be close enough to the EVC so as to merit only minor tweaks to the applications of the model to vigilance.

Speaking of which…

3.5 The Maintenance Dimension of Vigilance

In this section, I will outline the connections between the EVC model and vigilance. Roughly, the maintenance dimension of vigilance maps most directly onto the various components of the EVC model (though I’ll also argue that maintenance corresponds to the EVC, too, so that the EVC model provides a fully computational account of vigilance). This will then give us a template for discussing various failures of vigilance. In the next section, I’ll discuss the relationship between EVC, vigilance, and self-control.

In the Hot Car example, the parent plans to drop the kid off at school before heading to work, but ends up going straight to work. This is a present-directed case of maintenance failure. The parent fails to adequately respond to goal threats and goal substitution. What happens is that, in general, favoring task sets that minimize the cost of control leads to reliance on habitual routines. The parent, when leaving for work and putting the kid in the backseat, fails to accurately assess the need for engaging in controlled processing for goal maintenance. In one sense, this is understandable. As we
saw in the previous section, there is a tendency toward minimizing the use of control. Given the past record of success when relying on habits, there seems to be no need to deliberate about the use of control, much less initiate an actual deployment of control. From the parent’s perspective, both the likelihood of failure and the task demands are so low that there seems to be no need to think very hard about the task of dropping the kid off.

This can be represented in terms borrowed from the EVC model. Recall that there are three representational components to the model: the probability of achieving an outcome given a signal, the value of the outcome, and the cost of the signal. Misrepresentations could occur at each point. The parent might misrepresent the probability of achieving the outcome in the absence of a control signal, she might misrepresent some aspect of the value of the outcome (though this is unlikely), or she might misrepresent the cost of the control signal. As the case is described here, it seems that the parent misrepresents the cost of control. Rather than move from a default to controlled configuration, the parent instead sees no need to pay the cost of control.

This might seem fishy. You might think: “Wait, the parent is misrepresenting? Surely people don’t perform these computations at the personal level. These misrepresentations are subpersonal glitches, not anything you can pin on the agent.” The answer to this objection, however, provides a bridge between the agent and the computational principles outlined in the previous sections. Roughly, the computational failure can be attributed, in failure of vigilance cases, to the agent’s failure to construe her tasks appropriately (in other words, the agent fails to structure her plans appropriately, leaving those plans susceptible to goal threat or substitution).
The way that an agent construes her goals is important. Literature on implementation intentions suggests that concretely structured plans are less likely to be abandoned prematurely (cf. Gollwitzer, 1999). Similarly, task construal can have an impact on how attractive some tempting item seems (Fujita et al., 2006). Task construal bears on plan structure, as task construal typically implies a certain plan structure. For instance, there’s enormous difference between forming a plan to exercise as ‘I’ll workout when I have the time’ and ‘I’ll only take a rest if I’ve worked out the previous two days’ (Fujita, 2011). Similarly, recovering alcoholics who adopt a narrower, more present-directed focus on their circumstances (low-level construal) tend to relapse at higher rates than recovering alcoholics who adopt a more global perspective (high-level construal) (Keough, Zimbardo, & Boyd, 1999).

This might make it seem like all failures of vigilance are just failures of construal, thereby implying that there is just one kind of failure of vigilance. With the EVC in hand, we are in a position to outline more precisely four different kinds of vigilance failure. This account depends on the functional distinctness of maintenance and monitoring.

One speculative difference between monitoring and maintenance concerns the calculation of probability. Recall, from Section 3.4, that separate mechanisms calculate probability estimates for achieving an outcome. Perhaps the mechanism of monitoring consists in these probability calculators, whereas the mechanism of maintenance consists in cost calculators (value representations would be subserved by valuative—possibly reinforcement-learning—mechanisms). Hence, the aim of monitoring at the information-

---

18 There is some evidence that low-level construal sometimes benefits self-control (see Schmeichel et al., 2011).
processing level is probability calculation, while the aim of maintenance at the same level is cost calculation.

This connection to the EVC suggests the following taxonomy of vigilance failures:

<table>
<thead>
<tr>
<th>Functional Type</th>
<th>Temporal Aspect</th>
<th>Psychological Kind</th>
<th>Computational Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance</td>
<td>Present-directed</td>
<td>Failure to preserve task set</td>
<td>Inaccurate cost representation</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Future-directed</td>
<td>Coordination failure</td>
<td>Inaccurate future cost representation</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Present-directed</td>
<td>Failure to recall task set</td>
<td>Incomplete representation</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Future-directed</td>
<td>Double booking</td>
<td>Inaccurate probability representation</td>
</tr>
</tbody>
</table>

A coordination failure consists in inaccurately representing future costs of engaging control. In the paper deadline example, the individual misrepresents the balance of costs and benefits associated with engaging control, thereby committing to a task set that she likely will not carry through. In short, the agent fails to appreciate the task demands and the available resources (or lack thereof) in the future environment. Similarly, a failure to preserve a task set consists in inaccurate representation of the current costs of control.
Double booking consists in inaccurate probability representations. That is, when double booking, Mr. McIntosh fails to represent the future probability of achieving outcomes given controlled signaling. This is a little heavy-handed, as the future probability of successfully achieving inconsistent goals simultaneously is zero. But the example brings out what goes wrong in double booking.

The case of present-directed monitoring failures requires some discussion. Part of the activity of monitoring is to accurately represent features of the action context (either present or future) relevant to one’s planning. At the computational level, this amounts to accurately entering values for the various arguments in the EVC function. For example, suppose I’m telling a funny story around you that recounts your doing something embarrassing. I don’t even recognize that you’re getting upset and I continue telling the story. Here, I’ve misinterpreted the situation such that certain values (like your taking offense at the anecdote) are not even represented in my decision procedure to continue telling the story. This seems to be a failure of monitoring, but on a slightly different order than other kinds of failures of vigilance.

Note that, in the above, each aspect of vigilance maps to a component of the EVC. This strengthens the claim made earlier that the EVC provides a fully computational account of vigilance. Additionally, the EVC model leaves room for the importance of task construal. The representation of plan structure might play an important role at the computational level. However that might turn out, the foregoing at least suggests that failures of monitoring and maintenance are distinct, even though they both share a common constituent, namely task construal. This also demonstrates the dynamic interactions between the different dimensions of vigilance. For instance, the
representation of cost might be sensitive to representations of probability and vice versa (consider the Hot Car case; perhaps the representation of probabilities affects the representation of current costs of control. Not enough is currently known about the relations between these different aspects of the computations, so nothing definitive can be stated about dependence or fundamentality of one relative to the other).

The EVC model provides a computational underpinning for the functional account of vigilance. The reason for connecting these two constructs is not to reduce vigilance to the EVC model. Of course, an individual that realized a psychology that failed to be accurately described by either the computational or representational theory of mind might have vigilance. The point of tying the vigilance account to the EVC model is that it provides a computationally tractable way of thinking about the activity of vigilance. When tied to a sufficient mechanistic model of vigilance, EVC enables predictions about when and where vigilance will breakdown and how we might improve the exercise of vigilance.

3.6 Self-Control as Vigilance (II)

What remains is to determine whether the account of vigilance and cognitive control adequately captures certain features of the self-control construct. In this section, I’ll show how vigilance explains the activity of various processes typically associated with self-control. Additionally, I’ll consider two potential problems with the self-control as vigilance theory, namely that there are empirical reasons to dissociate cognitive control and self-control and that the function of self-control is exhausted by temptation resistance and impulse control.
There are three foundational components typically associated with models of self-control: inhibition, task updating, and task switching (see Davisson & Hoyle, 2017 for review). Additionally, there are four components typically associated with cognitive control: working memory, response selection, response inhibition, and task-set switching (e.g., Sabb et al., 2008). The EVC model makes explicit connection between cognitive control and vigilance. To connect this to components of self-control, consider that vigilance structures the flow of information into working memory. One aim of the activity of monitoring is to bring plan-relevant information to mind at the appropriate time, while filtering out information when it is no longer relevant. Bringing information to mind here amounts to bringing the information into working memory (Baddeley, 2012). This shows that updating is a component of vigilance and, by extension, that cognitive control is partly constitutive of self-control. As for inhibition and switching, these components fall under the maintenance dimensions of vigilance. Sending a control signal (which is constitutive of the activity of vigilance) is constitutively relevant to inhibiting some response for the sake of selecting some other response.19 Similarly, when the cost of some ongoing controlled activity becomes too great (relative to other goals that one could pursue), then the agent shifts. Again, this is just the function of the maintenance dimension of vigilance in its present-directed aspect. Thus, it seems possible to map both cognitive control and self-control onto the functional account of vigilance, thereby linking all three constructs.

19 See the discussion of constitutive relevance in Craver (2007).
One worry about this account is that it seems to turn self-control back into mere impulse control/temptation resistance. After all, if conflict monitoring is part of the maintenance dimension of vigilance (which dimension, recall, is the one associated with self-control), then isn’t self-control still just a form of conflict management? But the objection gets the view wrong. Conflict monitoring may involve detection of distractions, temptations, or impulses toward short-term rewards. However, note that ‘conflict’ applies to a much wider range of phenomena than this. Any creature with a heterogeneous set of goals that cannot be simultaneously pursued will need self-control. And the conflicts that get monitored might simply be conflicts between pursuing one or another goal at any particular time. Hence, there is a need for conflict monitoring even in the absence of being distracted or tempted.20

One final objection to consider concerns empirical evidence that self-control and cognitive control are distinct constructs. In a recent study, Scherbaum et al. (2018) proposed that self-control and cognitive control recruit distinct processes and are, therefore, distinct constructs. They hypothesized that if cognitive control and self-control rely on the same set of processes, then “experimental manipulations that cause control adjustments in one task should also increase controlled behavior in the other task” (2018: 195). The experimental design used two tasks, a cognitive control task (Simon task) and a self-control task (intertemporal value decision). Each trial pair consisted of the cognitive control task followed by the self-control task. Participants had two seconds to perform the

20 This contradicts Gary Watson’s assessment of the role that the virtue of self-control plays within the moral life: “Self-control is a virtue only for beings who are susceptible to motivation which is in potential conflict with their judgments of what is good to pursue” (1977: 322). Self-control, as a component of vigilance, is a virtue only for planning agents that have to manage pursuit of heterogeneous goods across time.
Simon task before immediately moving to the self-control task. Every participant completed 576 trial pairs. The results showed that there is “no difference in the probability of [choices exhibiting high self-control] for decisions following conflict Simon trials…and decisions following non-conflict Simon trials” (2018: 196). Hence, they conclude that self-control and cognitive control recruit distinct processes.

There are two problems with this, however. The first is a small point with experimental design. There is some evidence that when individuals perform a control-demanding task, they exhibit a tendency to continue performing the same task despite being cued to switch. This is known as the ‘task-set inertia’ effect (Allport et al., 1994; Yeung & Monsell, 2003; Yeung, 2010). The experimental design, however, did not rule out the possibility of task-set inertia effects. Hence, participants might have directed control resources toward the Simon task, relying on more habitual/default routines to complete the self-control task. The short time scale over which participants completed trials does not compensate for potential inertia effects.

The second problem concerns the hypothesis, namely that if cognitive control and self-control share similar processes, then tasks that activate cognitive control processes should elicit higher degrees of self-control in subsequent tasks. This, however, does not hold true even in cases where the two tasks are cognitive control tasks. For instance, Yeung et al. (2006) found that participants perform a cognitive control task more slowly and less accurately following the performance of a different cognitive control task when compared to performance of the same cognitive control task (there were also small ‘restart’ costs observed on same-same cognitive control task blocks). Hence, there is no reason to expect the claim that if two constructs share similar processes, then performing
one kind of task will boost performance of the other kind of task. For these reasons, we can dismiss the supposed evidence in favor of the distinction between cognitive control and self-control.

With this model in hand, we can explain some phenomena connected with traditional studies of self-control. The first concerns the relationship between self-control and effort. The underlying neurological basis of self-control, on this view, is the cognitive control system implicated in maintaining task sets. Allocations of control, on this view, are determined by cost-benefit analyses and so are partly sensitive to the costs of control. This provides a naturalistically plausible framework for thinking about the feeling of effort. Effort is the subjective correlate or phenomenological component associated with the costs of control, specifically the opportunity costs associated with allocating control in a particular way (see Cohen, 2017; Kool, Shenhav, Botvinick, 2017; Kurzban et al., 2013; Shenhav et al., 2017). However, this experience of effort need not always be aversive. This is because sometimes the lost opportunity costs are not valued more than the goals pursued in allocating control. When control is aversive, this might be due to diminished motivation or willingness to pay the costs of control (see Botvinick & Braver, 2015; Braver, 2015; Winecoff & Huettel, 2017). This would then incorporate a motivational element into the ‘self-control as vigilance theory’ (or, Vigilance Theory), thereby explaining the appeal of purely motivational views of self-control posited before.

Secondly, the view shows how impulse control/temptation resistance falls under the purview of self-control. Succumbing to temptation or giving into impulse is a kind of recovery failure. Typically, impulses and temptations have pull in virtue of being goal-relevant (though these goals might be maximally coarse-grained, like ‘experience
pleasure’; see Levy, 2011: 143). Hence, being in the presence of a temptation or under the pull of an impulse threatens to shift goal pursuit. Self-control is needed to recover from the initial pull of temptations and impulses. This recovery consists in reallocating control to manage conflicts between the goals activated by temptations/impulses and the goals one was originally pursuing.21

Ego depletion effects also make sense on the vigilance view. In general, the Vigilance Theory predicts that in a traditional two-task paradigm (like those used in original ego depletion studies) agents lose motivation to perform the task rather than run out of metabolic resources. In other words, the opportunity costs associated with allocating control to task performance becomes too high and agents shift either to focused thinking about something else or mind wandering (Seli et al., 2018-a). This connects to an old interpretation of the vigilance decrement (the deficit in performance associated with continued performance of a vigilance task), namely that as the stimuli associated with task performance loses novelty, the subject begins to think of other things (Broadbent, 1958). In other words, when agents get bored, they shift to other tasks and therefore perform worse.

This also answers some criticisms addressed to attention/motivation theories of self-control. For instance, Baumeister & Vohs (2016: 100) offer two criticisms:

21 There is some question as to what distinguishes temptations and impulses from ordinary conflicts that arise from plural goal pursuit. This, I think, is not a question that falls on the Vigilance Theory to answer, but I suspect the answer has something to do with agential autonomy. That is, something is a temptation or impulse in virtue of the fact that the goal pursuit activated by the temptation or impulse is either counter-resolutional (as in Holton, 2009) or that the agent does not identify or endorse the goal pursuit (as in cases of addiction relapse). This, however, has more to do with the metaphysics of autonomy and the psychology of commitments than with self-control.
[Some suggest] that fatigue in general has nothing to do with low energy but is instead a signal to interrupt one’s activities, reflecting opportunity and regulatory costs of perseverance. But what are those regulatory costs if not expenditure of energy?

...The opportunity cost argument has difficulty explaining the multi-task paradigm findings. If fatigue were merely a signal that it is generally a good idea to switch tasks (as opposed to being a signal that one’s energy has been somewhat depleted), why would it transfer so that fatigue from the first task is still felt during the unrelated second task (and indeed impairs performance on it?).

Consider these criticisms in reverse order. The second criticism we already considered by suggesting that subjects get bored and shift attention. The reason why this effect carries over between tasks is that subjects lose their motivation to perform in-lab activities (Nicholls et al., 2015). This also explains why increasing incentives on task performance between tasks in the ego depletion paradigm eliminates the depleting effect entirely (Muraven & Slessareva, 2003). When an agent sufficiently values performance on the second task, the opportunity costs reset and the agent performs close to ceiling (there are also reports of limited feelings of mental effort in these cases; see Boksem & Tops, 2008; Lorist et al., 2005).

The first criticism is simply a demand to understand what grounds the cost of control if not depletion of a metabolic resource. Recall from Section 3.3 that the ground of opportunity costs is the shared representational resources used in the brain (i.e., multiplexing). Because of the limited representational resources available for use in cognition, there are costs associated with allocating control (devoting representational resources to one processing pathway over another) that give rise to opportunity costs. Hence, the theory of self-control offered here can answer the criticisms raised against similar views.
3.7 Conclusion

This chapter showed three things. The first is that the standard story of self-control is incorrect. When we reorient our understanding of self-control around trying to understand limitations associated with control that derive from representational resources (rather than metabolic resources or brute structural limitations), a biologically plausible view of self-control emerges based on cost-benefit analyses of allocating control to a single task.

We also saw how this theory of self-control connects to vigilance. As vigilance manages goal maintenance, and self-control consists in allocating control to maintain goal pursuit, self-control is a species of vigilance (the two are not equivalent, as monitoring is not reducible to maintenance or self-control functions). There is a way to translate self-control functions into the language of vigilance when we characterize self-control functions in terms of maintenance. This also suggested a way to connect the subpersonal computational story about the allocation of control to personal-level properties. In particular, a crucial aspect of maintenance is task construal, and the representation of certain task sets at the computational level is sensitive to personal-level construal. Hence, vigilance (or the self-control aspect of vigilance) is not just a matter of subpersonal computations.

Finally, we saw how the vigilance theory of self-control explains a number of phenomena related to self-control such as effort and ego depletion. Additionally, the vigilance theory can supply answers to criticisms normally leveled against motivation/attention views of attention.
One theme of this project is that planning agency requires a distinctive kind of psychology. Acting over time is different from concatenated sequences of actions at a time. This shows that many elements of planning psychology are temporally modulated (or have future-directed aspects). From this perspective, one issue with the standard story of self-control is that it is built on the needs of agents acting at a time. Even discussions of resolutions and temporally extended wills are all about preparing to act at a time. A planning agent needs something different. This account of self-control takes the diachronic dimension of self-control as fundamental. This is reflected in the EVC computation containing an essential future-oriented component. This is part of a mechanism for future-directed self-control and, hence, fits the needs of a planning agent.
CHAPTER 4:
RESPONSIBILITY AND VIGILANCE

4.1 Introduction

We tend to think that others are responsible for what they do. When someone says that they haven’t done anything, this can typically function as an excuse for some alleged bit of wrongdoing. The notion that activity is the locus of responsibility makes lots of intuitive sense, as well. After all, if there’s something an agent does that brings about a bad or harmful state of affairs, there’s a direct line that exists between a positive feature of the agent and the wrongdoing.

For all its intuitive sense, however, there seem to be many cases where we hold others responsible for what they don’t do. Sure, some of these inactions (or omissions) can be motivated. Maybe somebody “forgets” that their brother-in-law asked for help moving and “unwittingly” ended up sleeping in. This kind of deliberate or willful

\footnote{My use of ‘responsibility’ refers to moral responsibility as opposed to causal, legal, or other forms of responsibility. The sense of moral responsibility at issue here is the accountability sense rather than the attributability and answerability senses of responsibility (cf. Shoemaker, 2015). A culpability-imputing form of blame, then, is central to the sense of moral responsibility that I discuss. I focus on moral blameworthiness (as opposed to praiseworthiness or creditworthiness), where moral blameworthiness picks out a range of judgments and attitudes bound up in non-trivial ways to the Strawsonian reactive attitudes of resentment, indignation, and guilt (Strawson, 1962). Lastly, I look only at responsibility for what one has done. I mention these distinctions here to forestall potential confusions with my use of the term ‘responsibility’.

111
ignorance still has the essence of action even if it has the veneer of omission. But other kinds of omission seem different. Consider the person who forgets to bring important materials for a work presentation after rushing out the door in the morning. Or the person who forgets to pack toothpaste for a work trip. Or the person who forgets to call their friend on her birthday. These failures are unwitting omissions, or failures to do something one ought to do where there is no recognition of one’s failure (until, perhaps, it’s too late).

People have recently taken interest in these cases of unwitting omission because of a strange tension in our thinking about such failures. On the one hand, unwitting omissions seem familiar and pervasive, and holding others responsible for such omissions seems deeply woven into our everyday practice of assigning and taking responsibility. On the other hand, unwitting omissions are paradigmatic cases of obliviously not doing anything. There appears to be no positive agential feature that can serve as the target of a responsibility ascription (see Amaya & Doris, 2015).

In this chapter, I connect the theory of vigilance developed in the previous chapters to the problem of responsibility for unwitting omissions. There has been no discussion yet of the normative dimensions of vigilance except for the brief sketch of the norms of vigilance in Chapter 1. However, there are rich normative applications of the theory. Failures of vigilance are a kind of unwitting omission. Hence, understanding how we are culpable for some failures of vigilance might provide a general template for thinking about responsibility for unwitting omissions.

The theory of responsibility offered in this chapter is not cut from wholly new cloth. I outline a kind of reasons-responsive theory of responsibility, in contrast to an
attributionist, quality of will, valuationist, virtue theoretic, or some other nonvolitional theory of responsibility. My sense is that nonvolitional theories of responsibility cannot accommodate our intuitions about responsibility for unwitting omissions. Making that argument, however, takes us well beyond the scope of a single chapter (for more on this, see Murray et al., 2019; Murray, Manuscript). Because of the large scope of the issue, not every view can be adequately considered. Developing one promising account, then, seems preferable to a hasty survey of an expansive literature.

Several reasons-responsive theorists have offered accounts of responsibility for unwitting omissions that are consistent with the general principles of reasons-responsive theories (Clarke, 2014; Murray, 2017; Murray & Vargas, Forthcoming; Rudy-Hiller, 2017; Vargas, Forthcoming). The common thread through these accounts is that there are legitimate norms that govern our awareness. These norms target agential capacities (reasons-responsive mechanisms), so failing to live up to these norms constitutes some form of wrongdoing. In some circumstances, there is no standard excuse available for the agent, and so the wrongdoing licenses blame.

Two recent objections have been raised against these accounts. The first is that the reasons-responsive account of responsibility is unfair. Roughly, this is because the capacities described in the aforementioned reasons-responsive accounts are subpersonal. Hence, when they fail to be properly exercised, this is simply a cognitive glitch and not something attributable to the agent (see Rudy-Hiller, Forthcoming; Rosen, 2015).

The second is that the reasons-responsive account cannot explain why we’re sometimes responsible for our *moral* ignorance. Some unwitting omissions occur because of factual ignorance (e.g., being ignorant of the fact that today is a friend’s birthday).
Other unwitting omissions occur because of moral ignorance, or a failure to apprehend some moral fact. For example, I may think it’s morally permissible to own slaves in light of my genuine ignorance that owning other persons is incompatible with the fundamental dignity that each person has. The problem here is particular to the version of the reasons-responsive account that I favor. On my view, responsibility for unwitting omissions is explained in terms of failures of appropriate vigilance, where these are failures to live up to certain demands of information management that derive from the social roles one occupies. However, moral ignorance seems importantly different from a case of mismanagement. With information mismanagement, an individual fails to recall some bit of previously known information at the right time. With moral ignorance, an individual fails to appreciate some fact that was (presumably) never known in the first place. So the vigilance-centered theory of responsibility seems inadequate to the task of explaining responsibility for moral ignorance.

One common factor for both problems is that each is connected to a social dimension of agency. Issues of fairness concern relations to others and the legitimacy of how others exercise authority over us. And even hard-line moral realists can concede that moral knowledge is socially and culturally mediated. This suggests that careful attention to socially scaffolded elements of agency might help to fill out the reasons-responsive account of responsibility for unwitting omissions. For this reason, I consider the social dimensions of vigilance (in Section 4.4), focusing on the way in which one’s roles and desire to exhibit social competence structure vigilance. This, in turn, informs my discussion of blame and fairness (in Section 4.5) and responsibility for moral ignorance (in Section 4.6).
4.2 The Basics of the Reasons-Responsive Theory

A standard reasons-responsive theory of responsibility states that an agent is responsible for some action or outcome when the mechanism that mediates the action (or is causally implicated in bringing about the outcome) is suitably reasons-responsive. Details that distinguish different reasons-responsive theories turn on what constitutes a suitable degree of reasons-responsiveness, the conditions of mechanism ownership, and the kinds of reasons toward which these mechanisms must be sensitive (Fischer, 2012; Fischer & Ravizza, 1998; McKenna, 2013; Stump, 1996).

These debates are orthogonal to the main concerns of this chapter. The important thing is that reasons-responsiveness is a central component of responsible agency. There is no need to hold that having a reasons-responsive mechanism is sufficient for being a responsible agent. It’s enough for present purposes that having a reasons-responsive mechanism is necessary for being a responsible agent.²³

A reasons-responsive theory of responsibility locates responsible agency in the possession and exercise of agential capacities implicated in the production of effective self-governed action.²⁴ Normally, reasons-responsive theories countenance two kinds of capacities as constitutive of this morally significant self-governance.²⁵ The first are those capacities that enable agents to engage in goal-directed behavior. Among these are the capacity for behavior regulation, the ability to adopt and update goals, and the

---

²³ Other conditions might require that certain kinds of indeterminism obtain in the causal nexus of agent, action, and environment. Debates about the compatibility of determinism and responsibility (or the claim that indeterminism is necessary for any kind of responsible agency) go well beyond the scope of the present chapter. For that reason, I won’t focus on these issues here.

²⁴ Cf. Vargas (2013: 203). Here, responsible agency just is the possession of capacities constitutive of effective and rational self-governance.

²⁵ Vargas (2013: 200). Some of my terminology diverges from that of Vargas.
coordination of plans over time. The other kinds of capacities constitutive of responsible agency are those that afford agents awareness of their moral environment (foresight, means-end reasoning, etc.). Thus, reasons-responsiveness is a feature of agents that exhibit suitable degrees of self-governance and awareness.26

Explanations of responsibility in terms of reasons-responsiveness pick out some suitable relation (at the moment of wrongdoing) between an agent’s responsibility-relevant capacities that are constitutive of reasons-responsiveness and some bad action or outcome as a necessary condition for ascribing responsibility for that action or outcome.27 These explanations are attractive, as mentioned before, because they pick out a relation that ties the agent to her wrongdoing in a way that makes it appropriate to blame the agent for her wrong conduct (see Sher, 2009: 73-75).

4.3 The Epistemic Condition on Responsibility

One attractive feature of the reasons-responsiveness theory of responsible agency is that the theory straightforwardly implies relatively simple conditions on responsibility. To adopt somewhat standard terminology, these are the Control condition and the Epistemic condition. Roughly, the Control condition states that it is appropriate to blame an agent for A-ing only if the agent has some reasons-responsive mechanism that mediates the agent’s A-ing. The Epistemic condition states that it is appropriate to blame

---

26 There is no need to suppose that there is a single, domain-general capacity constitutive of self-governance and another domain-general capacity constitutive of awareness. Instead, there is likely a patchwork of various capacities with different functional profiles across contexts. These various capacities would be tuned to a variety of different kinds of considerations (moral, prudential, factual, social, etc.).

27 In what follows, I will focus exclusively on behavior that seems to merit distinctively negative moral reactions.
an agent for $A$-ing only if the agent appreciates some suitable subset of the moral qualities of $A$-ing.

In short, an agent’s responsibility for some action is a function of having control over the action and knowing what she is doing in acting.

The Epistemic condition raises a problem for explaining responsibility for unwitting omissions on the reasons-responsive account. By definition, unwitting omissions are done without any awareness of the moral qualities of one’s behavior. Even further, for most unwitting omissions it is likely the case that there was no prior moment where the agent suspects that she will unwittingly omit or is at risk of doing so (in other words, there is little reason to think that every unwitting omission is a motivated failure). This generates a problem because without awareness, it seems like there are no legitimate grounds for blaming others for their unwitting omissions (Rosen, 2004).

The standard move for reasons-responsive theorists in light of this problem is to amend the Epistemic condition to include a capacity clause. That is, having awareness is not centrally important to responsibility. Instead, having the capacity for awareness is essential. This undergirds the claim that agents are sometimes responsible for their unwitting omissions because they should have known better (Clarke, 2014: 167; Lucas, 1993: 52; Murray, 2017: 521; Sverdlik, 1993: 141). To say that someone should know better is to implicitly presume the capacity to know better (Hart, 1968: 148).

The standard revision to the reasons-responsiveness theory in light of the problem of unwitting omissions, then, is to recognize that what matters to responsibility is not having specific occurrent mental states when acting wrongly; rather, what matters is having a certain kind of competence or capacity to navigate an action domain. That
competence may go unexercised (perhaps even obliviously so), but the reasons-responsive theorist need not automatically concede that there can be no responsibility in such circumstances.

The theory of vigilance developed in the previous three chapters describes a substantive, empirically tractable capacity that might underwrite some ascriptions of responsibility for unwitting omissions. This makes sense in part because paradigmatic failures of vigilance all seem to be cases of unwitting omission. Consider the cases mentioned in Chapter 2. Mr. McIntosh should have been more aware of the schedule. Parents that forget their children in the car fail to be vigilant in virtue of failing to be appropriately aware of their environment. Failures of vigilance appear to be (or result in) unwitting omissions. Hence, understanding vigilance contributes to understanding responsibility for unwitting omissions.

The capacities for control and awareness (including vigilance) are not sufficient on some occasion to hold some individual responsible for an occurrence. So, even though possessing vigilance is an important component of being responsible for unwitting omissions, there are additional conditions that account for when it is reasonable to expect someone to exercise these capacities. Section 4.4 proposes an account of these conditions.

4.4 Competence and Control

When people fail to be vigilant, presumably there is a story to be told about these failures, a story that reflects the operations of neurological, cognitive, and affective processes. But why view the outcome of those things—things that are, let us suppose,
determinative in failures of vigilance—as licensing criticism of unwitting wrongdoers? A satisfactory account must specify why the demands implicitly encoded in our responsibility ascriptions of unwitting wrongdoers are reasonable. An answer is forthcoming if we focus on the value of self-control.28

Many philosophers have noted that control or self-governance is central to responsible agency (Ekstrom, 2000; Roskies, 2010; Fischer & Ravizza, 1998; Vargas, 2013). On many such accounts, what makes an individual a morally responsible agent is (at least in part) the general capacity to direct oneself in light of distinctively moral reasons for acting. Perhaps the standard way to flesh out this general capacity is in terms of an agent’s (or some mechanism’s) responsiveness to reasons (Fischer and Ravizza, 1998; Nelkin, 2011; Roskies, 2012; McKenna, 2013; Vargas, 2013). According to these accounts, reasons-responsiveness is at least necessary (if not sufficient) for responsible agency. But why is control important in the domain of responsible agency?

Put simply, control matters because it entails the ability to conform one’s behavior to norms that arise as a result of one’s descriptive and aspirational self-conception. Here, a descriptive self-conception includes information that represents who one is, the role (or roles) that one currently occupies, and facts about one’s current social status. An aspirational self-conception includes information about what one hopes or aspires to be and the roles that one aims to occupy. In cases where one’s descriptive and aspirational self-conceptions overlap to some degree, two things are true: one has certain competencies that (at least partly) constitute one’s being that person and one reliably  

28 What follows in this section draws heavily on material from Murray & Vargas (Forthcoming: Section 5).
manifests those competencies in one’s behavior. So, if one is (and wants to be) a loving parent, one has the ability to be sensitive to the needs of one’s child and the wellbeing of that child and one reliably manifests or acts on those abilities. Control matters because it reflects the extent to which one actively seeks to, and succeeds in, conforming behavior to the self that one aspires to be, and thus (to some extent) is. The remainder of this section expands on this account of the relationship between self-conception, competence, reliability, and control.

One’s aspirational self-conception includes elements such as the relationships that one deems important, the projects that one finds fulfilling, and the activities that one esteems (hereafter, ‘self-conception’ functions as shorthand for ‘aspirational self-conception’). Despite localized or temporary fluctuations, in the ordinary case, one’s self-conception at a time (or over a minimal range of times) expresses many of the values, carings, and aspirations of a given agent at that time (or across that range of times).

Part of the content of one’s self-conception is an awareness of special skills or capacities that are necessary to manifest one’s overall scheme of cares. These capacities are important in at least three ways (cf. Schapiro, 2001). First, they play an important relational function. Our distinctive capacities relate us to the world in a way that reflects the kind of person we aim to be. Second, they play an instrumental role, in that exercises of these capacities bring about the activities that we esteem and realize the projects that

---

*While the descriptive portions of self-conception are interesting, the aspects of one’s self-conception that are important for this paper are the aspirational elements. Also, the notion of “self” at issue here likely varies cross-culturally, as Heine, Lehman, Markus & Kitayama (1999) suggest, but I will leave aside consideration of these issues.*
we deem fulfilling. Third, our capacities play a revelatory role, in that exercising our capacities reveals to others certain features of our self-conception. As such, our capacities play a crucial role in helping us live up to and realize our aspirations and ambitions.

Such capacities define a sphere of competence that ranges over a set of actions in which we can reliably engage.\textsuperscript{30} Thus, in virtue of one’s self-conception one takes oneself to have certain capacities that one can reliably exercise to bring about certain outcomes. If, for instance, part of one’s self-conception includes being a reliable teacher or a reliable parent, then some of the act-types that fall within one’s sphere of competence will include activities that fall under those roles.

Competence, here, is a function of the degree of reliability that one manifests in undertaking certain activities. Thus, there is an important relationship between self-conception, competence, and reliability. The central connection is that we want to display reliability in those domains where we value competence (where it fits into our ideal self-conception). Our valuing such competence also signals to others that we desire to be assessed in light of the norms that attach to occupying particular roles, the expectations that these norms generate, and the demands from others that they invite.

When some individual omits unwittingly, the individual violates demands that govern her in virtue of the particular roles that the agent occupies. Where an agent regards the demands as unreasonable, the demands can be protested. Plausibly, demands

\textsuperscript{30} Amaya & Doris (2015: 258-260) also discuss an agent’s “zone of secure competence” with respect to similar cases (what I call ‘unwitting omissions’ they call ‘performance mistakes’; as far as I can tell, performance mistakes appear to be a species of unwitting omission). However, Amaya & Doris claim that normative competences constitute the zone of secure competence, and equate these normative competences with appropriately reasons-responsive mechanisms. This discussion of competence provides a fuller characterization of what these normative competences are, how they link up with other aspects of our psychology, and why these competences are considered normative.
are unreasonable in at least two circumstances. The first is when the agent defensibly resists occupying that role. If you are no good in the kitchen, then your partner cannot expect you to cook a perfect soufflé. If they do have that expectation and subsequently blame you for burned dessert, you can reasonably object to being placed in that role. The second is when one’s circumstances contain systematic, competence-defeating features. Suppose you are a good cook, but you’re preparing a dish in someone else’s (vastly underequipped) kitchen, where the oven has a faulty temperature gauge. When the soufflé comes out of the oven burnt to a crisp, criticism is unreasonable because the circumstances conspired to thwart the exercise of one’s competencies. Other things equal, the unreasonableness of a demand in either sort of circumstance gives me an excuse for certain kinds of failure.\footnote{Of course, there will be some question as to how we sort out the excusing kinds of ignorance from the non-excusing kinds, though some of that machinery will derive from a more systematic account of the norms of vigilance. In spite of this problem, it remains true that there are relatively clear-cut examples of ignorance that make a demand unreasonable and thereby furnishes the agent with an excuse.}

Where the demands are reasonable, however, there are two primary responses to mistakes that are available. Either the community must adjust its expectations of the agent (e.g., I am not a good cook after all) or the agent merits blame (Higgins, 1987). There is oftentimes a high social cost to signaling to others that one lacks the relevant competence. So, individuals are ordinarily motivated to live up to these expectations, and to accept negative assessments in light of failures to meet these expectations (Higgins, 1987; Miller & Monin, 2016). Accepting a negative assessment is an important signal to the community. In a costly fashion, it communicates that one takes oneself to have the
relevant competence that figures in the demand, even if one did not (or could not) successfully exercise the capacity in the criticized case (cf. Sinnott-Armstrong, 1985).

This helps us to see the personal significance of competence. Because agents desire to exhibit competence, individuals have an interest in manifesting reliability in those domains they find valuable. Put another way, agents have an interest in manifesting their reliability to themselves and others for the sake of demonstrating that they can be considered reliable. When individuals omit unwittingly in domains of personal importance, there is a *prima facie* threat to that individual’s self-conception in virtue of the threat to that individual’s claim to reliability in a particular domain. One way to defuse this threat is to take responsibility for mistakes. In taking responsibility (either moral or non-moral) for some mistake, an individual signals both to herself and others that she possesses the relevant competence (Crocker & Wolfe, 2001). Acknowledging a failure to exercise this competence, even when it is seemingly costly, nevertheless affirms the agent’s competence in that domain. Hence, the sphere of competence helps to explain why individuals would agree to take responsibility for their mistakes.

Importantly, mistakes only threaten one’s self-conception (and thus become something for which an agent might take responsibility) when the mistake threatens the claim to reliability. If an agent already has a solid reputation for reliability, then individual mistakes will not constitute a threat. For example, if Christen Press makes a bad run or LeBron James throws a bad pass, nobody will question their competence as professional athletes. Systematic displays of unreliability or relatively grave singular lapses, however, will threaten even the maximally competent agent. At bottom, then, the
agent’s claim to reliability is the foundation for the normative account of expectations and offered here.

An agent’s sphere of competence is crucial for control. One possesses self-control to the degree that one’s behavior falls within one’s sphere of competence. By extension, one counts as a responsible agent only to the degree that one is acting within one’s sphere of competence. This implies (virtuously) that self-control and responsible agency are highly context-sensitive and shift according to one’s movement into and out of one’s sphere of competence.

All of this supports the contention that control plays a valuable role in our lives, which, in turn, explains the centrality of control in reasons-responsive theories of responsible agency. These considerations support a normatively significant account of control in terms of the possession of capacities that figure into our self-conception and constitute a sphere of competence. On this account, we are in control of ourselves and what we do simply in virtue of possessing these capacities in a context where nothing blocks or significantly impedes the proper exercise of these capacities.  

In this way, the competence account explains why certain demands for vigilance are reasonable. It also explains why agents accept these demands and accept criticism when they fall short of these demands (absent the possession of some excuse). And this

---

While the discussion of competence is indebted to Raz’s (2011) account, it departs from him in significant ways. One notable way is that Raz’s account of competence focuses entirely on the agent’s beliefs about her competence and self-conception. Some have criticized Raz’s account on just this point (see, e.g., Watson, 2016) since it seems that agents can be and are deceived (perhaps systematically) about the abilities that they have and the degree to which agents can reliably exercise these abilities (see Kruger & Dunning, 1999). This account, however, avoids this criticism. While it incorporates agential beliefs into the story, de facto reliability is more fundamental. While beliefs about one’s abilities and competence have some downward influence on de facto reliability, empirical research suggests that beliefs cannot account for the full range of facts about agential competence and reliability. In this way, this account of competence is better suited to the empirical facts than Raz’s account.
account reflects the context- and interest-sensitivity of these norms, suggesting that what counts as a reasonable demand varies greatly across a variety of contexts and between various individuals.

The account offered here has one notable upshot. The emendation to the reasons-responsive theory of responsibility required some story about a capacity that the agent should have, but failed to, exercise in omitting unwittingly. The competence account specifies what is meant in saying that unwitting wrongdoers should have known better, and the account of vigilance offered throughout these chapters supplies a substantive account of the relevant capacity.

4.5 Fairness and Cognitive Glitches

On the account developed here, an agent is responsible for her unwitting omission when the unwitting omission manifests a lack of appropriate vigilance. A lack of appropriate vigilance is one way of manifesting sub-optimal control in one’s behavior, so that explanations of responsibility that refer to sub-optimal vigilance are continuous with reasons-responsive explanations of responsibility.

One possible problem with this account mentioned in the Introduction concerns whether failures of vigilance are the kinds of things for which agents should be held responsible. Given the argument of Chapter 3, the cognitive control system subserves the function of vigilance. But the cognitive control system is an organized collection of neurobiological processes performing various computations. A failure of vigilance, on this view, amounts to failures at this mechanistic level. Plausibly, people lack control over what happens at the neurobiological level. So, people lack control over whether
failures of vigilance occur. If people lack control over whether failures of vigilance occur, then failures of vigilance appear to be the upshots of cognitive glitches, something that happen to people from time to time.

We can make this problem with the account vivid through example. Suppose you enter a reminder on your phone about your upcoming anniversary. You then trust that, when the time is right, the alert will remind you to do some planning (e.g., book a reservation or purchase some flowers). However, between the time you set the reminder and the time the alert was supposed to go off, your phone went through a series of automatic updates (mercifully performed at 3am while you are sleeping peacefully). These updates, unbeknownst to you, happen to erase your alerts. The window for preparing adequately for the anniversary comes and goes, and you’re left to suffer the fall-out, pleading with your partner to understand that it is the phone’s fault! I swear that none of this is autobiographical.\(^{33}\)

In any case, this seems to describe a failure of vigilance. An agent adopts a plan and, given cognitive constraints, ‘hands off’ the plan to memory. While memory is typically reliable, it can also fail from time to time for various reasons. Those failures are things that happen and are not attributable to the agent (hence, not something for which the agent is responsible).

Following on this, assigning responsibility to someone for an unwitting omission seems somewhat unfair. If failures of vigilance are analogous to (or just are) a kind of cognitive glitch, then how can it make sense to hold somebody responsible for such a

\(^{33}\) Thanks to Gideon Rosen for suggesting this example. He assures me that the example would not form part of his autobiography either.
thing? It seems that, in cases of unwitting omission, what an agent does (or fails to do) is disconnected from her rational powers of self-determination and practical reasoning. If that’s the case, then what justifies blaming someone for an unwitting omission?

To say that failures of vigilance seem like cognitive glitches because an agent lacks control over the activity of subpersonal processes appears to be a category mistake. Recall that the cognitive control system, explicated in Chapter 3, is a constituent of vigilance, and vigilance itself is a constituent of an agent’s global capacity for effective self-governance. Hence, demanding that an agent control the components of her vigilance is nothing more than a demand that an agent control her control.

I suspect that the root of the objection is suspicion about the prospects of a reductive, componential analysis of agency. I have assumed, throughout this project, that conceptual accounts of agency proceed by analyzing the sub-components of larger agential capacities and explaining the structural and organizational properties of the underlying system of sub-components. Hence, agency consists in a structured organization of monitoring, memory, planning, valuative, and regulative systems (among other components) to contribute to manifestations of agency at the personal level. On this view, vigilance is a kind of executive supervisory mechanism centrally ‘located’ in the agent’s psychology. Assuming some kind of reductive account of agency, then, the claim that the agent ‘hands off’ plans to other components of her psychology makes no sense. The whole system, including vigilance, comprises the agent’s global capacity for effective self-governance. Of course, this raises the question of whether one should assume a reductive view of agency. However, I feel comfortable waving my hands here. We all have to adopt some assumptions (I’ve also assumed realism about persons), and in
the Information Age brand managers tell you that the most important thing is playing to your niche.

What, then, does this tell us about fairness? In general, questions about the fairness of responsibility attributions are thorny (see Rosen, 2015). The fairness response requires two steps. The first is to see that questions of fairness are not so much about the fairness of individual blame tokenings (e.g., whether it is fair for me right now to blame you for what you just did), but rather about the fairness of rules that organize and govern practices of assigning blame (Vargas, 2013). This reveals that the question is not about deserved blame, but whether the principle that justifies assigning responsibility to someone is fair (this, in turn, would then tells us whether blame tokenings are fair, because their fairness is entirely derived from whether or not blaming on this occasion is permissible under the various rules that govern the practice of blaming).

The fairness of particular principles is a function of whether or not the principle survives contractualist procedures for rule adoption. The procedure is a kind of cost/benefit analysis. Participating in practices of responsibility assessment carry with it certain risks. As mentioned in the previous section, being the target of various reactive attitudes is costly and implies the risk of losing social status. However, there are many benefits. Participating in a responsibility system is a way of managing one’s social status and reputational capital and provides a mechanism for signaling reliability to members of one’s community.

The goal here is not to determine whether there are benefits associated with the social practice of responsibility that outweigh the risks associated with licensing others to hold you responsible. We can assume that there are significant enough gains associated
with the practice of responsibility. The question is whether the principle that justifies assigning responsibility to unwitting omitters is something we should include within our practice. Recall that the principle states that we are in control of ourselves and what we do simply in virtue of possessing certain capacities in a context where nothing blocks or significantly impedes the proper exercise of these capacities. This principle about the nature of control then justifies holding others responsible for unwitting omissions (because, on this account, unwitting omissions are a manifestation of an agent’s admittedly sub-optimal control).

One proposal about the rules that structure our responsibility practices seems to rule out this principle. The overarching rule is that the scope of responsibility extends only to those things that are up to you in the sense that one’s behavior and attitudes are dependent on the desires, values, and character that one currently possesses (Lenman, 2006: 23). What is important to this rule is that an agent sees her decisions as dependent on her desires, values, and character, and accepting the rule is a reflection of seeing oneself as an effectively self-governing agent (Lenman, 2006: 24).

This might seem problematic for failures of vigilance and the associated principle about the kind of control relevant to responsibility. If forgetting is the sort of wrongdoing that is not connected to character, desires, or value (as I argued in Section 4.3), then why would anyone be motivated to enter a system that licenses sanctioning forgetfulness? Indeed, insofar as everyone knows they’re bound to forget something morally significant at some point, it seems we have lots of reason to resist entering such a system.

Consider, however, an alternative perspective on the issue. Resistance to entering such a system might betray a certain lack of self-trust. That is, one motivation for
accepting the relevant principle is that doing so reflects seeing oneself as a trustworthy agent. While that kind of self-trust might be a necessary condition on temporally-extended agency (Hinchman, 2003), it might also be a necessary condition on important forms of social agency necessary for joint activity (see Bratman, 2014: 89). Not willing to accept responsibility for forgetfulness (or, not willing to participate in a system that licenses sanctioning for failures of vigilance) manifests a lack of self-trust detrimental to one’s agency. As outlined in the previous section, our sense of self demands manifesting reliability to others to signal willingness and fitness for social engagement. In attempting to elide acceptance of the principle, one exhibits substandard reliability and an inability to get along with others.

Hence, the benefits of accepting the principle that justifies assigning responsibility for failures of vigilance include securing a desirable kind of respect linked to one’s reliability at occupying and fulfilling one’s social roles. And, while one also risks being held accountable for failures of vigilance, winning respect and manifesting self-trust seem to outweigh these costs. This does not mean that all failures of vigilance are culpable. The principle that justifies assigning blame for failures of vigilance itself interacts with other norms of blameworthiness, particularly norms that dictate exculpatory circumstances and mitigating factors.

4.6 Vigilance and Moral Ignorance

At this stage, the control-based theory of responsibility, supplemented with an account of vigilance, is able to explain culpability of agents whose unwitting omissions make them liable to blame. There is an account of control that suggests principles that
justify these attributions of blame, an account of why we accept demands for particular levels of vigilance, why these demands (and the corresponding principles that justify these demands) are fair, and a substantive account of the capacity that these demands target.

Failures of vigilance seem to manifest (or just be) some form of information mismanagement. There is an intention that figures into some plan, the intention is stored in memory, but a subsequent failure of vigilance results in one failing to recall the intention. This is information mismanagement, in a broad sense, because one fails to appropriately move information in and out of mind as is appropriate relative to one’s plans.

While the account outlined above explains responsibility for ignorance in cases where ignorance amounts to information mismanagement, the account seems not to explain an important class of cases that has recently been of interest to responsibility theorists. These cases describe agents that do not mismanage information, but seem instead to be totally unaware of the demands of morality. So-called moral ignorance seems culpable even though instances of moral ignorance are not necessarily also instances of information mismanagement. If this is the case, then it means the vigilance-centered theory of responsibility for ignorance offered here is limited to a certain range of cases. While a disjunctive theory of responsibility for ignorance might ultimately be correct, we might hope for the vigilance-centered theory to deliver a unified verdict on the conditions of responsibility for ignorance.

In fact, the situation is slightly more complicated. We can explain some cases of responsibility for moral ignorance within an abstract characterization of the reasons-
responsiveness theory itself without descending into the nuances of vigilance. To see this, consider two examples of moral ignorance from Gideon Rosen:

Smith is a run-of-the-mill American sexist circa (say) 1952. Like any decent middle class father he has encouraged his sons to go on to college, setting aside money for the purpose. But like any run-of-the-mill sexist he has done nothing comparable for his daughters. This differential treatment is not malicious. But it is unfair and therefore wrong. But of course Smith doesn’t know this. He doesn’t know that his daughters deserve equal consideration in this respect (2003: 66).

Ancient Slavery. In the ancient Near East in the Biblical period the legitimacy of chattel slavery was simply taken for granted. No one denied that it was bad to be a slave, just as it is bad to be sick or deformed. The evidence suggests, however, that until quite late in antiquity it never occurred to anyone to object to slavery on grounds of moral or religious principle. So consider an ordinary Hittite lord. He buys and sells human beings, forces labour without compensation, and separates families to suit his purposes. Needless to say, what he does is wrong. The landlord is not entitled to do these things. But of course he thinks he is (2003: 64).

Rosen’s purpose for introducing these cases is to provide examples of blameless moral ignorance. While I’ve omitted some details of the cases, it seems safe to assume that the individuals in these stories are not obviously blameless (we should be careful here; I’m concerned with responsibility for behaviors, such as beating one’s slave and not putting money into your daughter’s college fund; that means I’m not centrally focused on whether the individuals are responsible for the beliefs that are causally related to these behaviors). Safer still is the claim that intuitions are mixed on these cases (Faraci & Shoemaker, 2014; Kissinger-Knox, Aragon, and Mizrahi, 2018). Is there any way to introduce some order?

These reactions to the cases can be explained with the abstract theoretical machinery of reasons-responsiveness. Recall that, on the reasons-responsive theory, responsible agency is a matter of possessing certain capacities constitutive of effective
self-governance, and responsibility for something is a function of the relation that obtains between these capacities and the relevant something. Hence, capacity possession is central to responsibility. In the Ancient Slavery case, the Hittite lord is not responsible for beating his slave. One condition on capacity possession is that one possesses the concepts relevant to exercising the capacity. This seems true in cases where the capacities are conduct-controlling. So, if one lacks the concept of social hierarchy, the person is unable to govern herself in light of certain facts about status because such facts are inscrutable to the person. The same seems to be true of the Hittite lord. As Rosen expands on the case, he adds:

> Unlike race slavery in the Americas, ancient Near Eastern slavery was not supported by myths about the biological or psychological inferiority of the slave. One became a slave through bad luck or imprudence; in principle the status could befall almost anyone. It is less clear to what extent this ignorance was grounded in false religion. The evidence suggests, however, that there was no perceived need for theological rationalization. The institution of chattel slavery was simply taken for granted (emphasis original; 2003: 65).

This mixture of factual ignorance and contingent facts about location in the history of ideas results in people lacking the concept of, say, Equal Respect or Equal Dignity. In fact, it seems likely that such concepts did not enter public consciousness until quite late in human history. If this is the case, then these people—including the Hittite lord—lack the concepts relevant to moral deliberation about how to treat slaves (they also lack the concepts relevant to determining that being in a position where one needs to deliberate about slave treatment is itself morally wrong). Lacking the relevant concepts, however, amounts to lacking the relevant conduct-controlling capacity. The reasons-responsive theory entails that when an agent lacks a capacity relevant to controlling some behavior
(or bringing about some state of affairs), then the agent is not responsible for the occurrence of that behavior (or the state of affairs obtaining). At the very least, the reasons-responsive theory turns the question of whether the Hittite lord is responsible for beating his slave into an empirical question. That is, holding fixed other aspects of the case, did the Hittite lord possess the concepts relevant to the capacities needed to control his conduct appropriately? Turning philosophical questions into empirical ones is at least some form of progress.

The same explanation does not apply, however, to the sexist father. By 1952, modern Western civilization had incorporated the idea of respect for persons and the equal dignity of all human beings into the collective moral consciousness (in theory, if not in practice). So any appeal to lacking the relevant concepts holds no water. However, the reasons-responsive theory entails that a thwarted capacity provides an excuse to the agent. And the sexist father seems to be in an environment that thwarts his capacity to appreciate the demands of respect and equal dignity. Here’s Rosen again:

Smith is the sort of complacent sexist who takes it for granted that his sons have legitimate expectations to which his daughters are not entitled (and perhaps vice versa). Let’s suppose in addition that his commitment is not based on some sort of theory—some bit of bad religion or bad science. Let’s assume, in other words—and this is hardly unrealistic—that Smith believes what he believes because he finds it obvious, and that he finds it obvious because he was raised to find it obvious and because the people he takes seriously find it obvious. The idea that gender matters in this way thus functions for him as an undefended axiom of moral common sense (2003: 67).

Smith’s environment is such that his capacity for appreciating the demands of respect and equal dignity are thwarted (reinforced by the fact that everyone around Smith and everyone that Smith takes seriously share, at least at some level, Smith’s moral outlook).
We can make the case for a thwarted capacity more acute by noting an epistemic analogue to the case. Consider a child, George, whose parents are thoroughly convinced that the earth is flat. This item is a central component in the parents’ folk cosmology and they do their level best to convince George of the truth of the belief. From the earliest years the child is told that supposed photographs of the Earth from space, which appear to show that the Earth is spherical, are simply carefully crafted government deceptions. All such evidence, they are told, must be reinterpreted in light of their knowledge of the conspiracy. In addition to George’s parents, his neighbors, friends, church leaders, teachers, and even his favorite NBA player all adhere to the belief that the Earth is flat. In this case, is George criticizably irrational for his belief that the Earth is flat? The answer seems to be no.

Notice, however, that the case exactly parallels that of Smith, with the exception that Smith holds a false moral belief, while George holds a false factual belief. The environmental factors, including the widespread pervasiveness of the belief in question, conspire against both George and Smith to thwart their arriving at true conclusions about the world (in this sense, one difference between George and Smith is that George’s false belief doesn’t seem to affect his conduct-controlling capacities, while Smith’s false belief does preclude him from rationally controlling his conduct in light of other, general beliefs that he holds).

Now, you might be thinking this is rather odd. It seems like the reasons-responsive theory is a way of getting people off the hook for their moral ignorance. Maybe you’re with me in thinking that sometimes moral ignorance is blameless, and that ancient slaveholders and complacent mid-century middle America sexists aren’t always
liable to blame for the behaviors that follow from this aspect of their moral outlook. But this doesn’t hold generally—right? After all, aren’t there some cases where people are on the hook for their moral ignorance?

In those cases where the answer is ‘yes’, we seem to be back into the problem that the vigilance-centered aspects of the reasons-responsive theory do not seem explanatorily relevant to cases of moral ignorance. Above, we considered one way that Smith’s sexism might be excused on the reasons-responsive theory of responsible agency. But suppose that we think it’s a case where Smith is liable to blame. How might this be explained as a failure of vigilance?

The problem is that cases of moral ignorance, as noted before, do not seem to be instances of information mismanagement. Neither the Hittite lord nor Smith the sexist seem to fail at recalling some plan or implementing an intention at the appropriate time; rather, their problem is that both failed to be aware at all of information relevant to controlling their conduct in accordance with the demands of morality.

Here are two speculative possibilities for explaining these cases of moral ignorance in terms of failures of vigilance. At one level, failures of vigilance exhibit a kind of information mismanagement. However, we can abstract to a higher level of generality in characterizing vigilance and failures of vigilance. Exercises of vigilance consist in noticing and appreciating certain considerations. This requires allocating attention properly and being disposed to recall these considerations when relevant. All of this is a matter of properly allocating one’s psychological resources over time. Described in this way, a failure of vigilance amounts to improper allocation of one’s psychological resources. This results in two possibilities for explaining moral ignorance as failures of
vigilance. First, it may be the case that being human is such that it entails occupying certain roles. Some considerations relevant to inhabiting these roles are never excusably missed. Second, it may be that some considerations are so obvious that missing them can only be achieved through mismanaging one’s psychological resources over time.

I am not convinced that these two possibilities are mutually exclusive. Putting cards on the table, I am skeptical that there are any such things as universal human roles, but those inclined to moral psychologies with a Kantian flavor might find this idea more attractive. I do think that the second aspect enjoys the advantage of explaining the reactions of people to cases of moral ignorance. The wrongness of sexism and chattel slavery are just too obvious for anyone to be innocently unaware of these moral qualities. This unawareness, as a consequence of failing to appropriately manage psychological resources, is itself a result of failing to be vigilant. In this way, we are able to connect responsibility for moral ignorance to the vigilance-centered aspects of the reasons-responsive theory of responsibility offered here.

4.7 Conclusion

In this chapter, I discussed some of the normative applications of the theory of vigilance developed so far. One obvious normative application is to explain responsibility for unwitting omissions. This is because paradigmatic failures of vigilance are all unwitting omissions (or result in unwitting omissions). Explaining responsibility in such cases has been difficult on a reasons-responsive theory of responsibility. However, including elements of the vigilance theory provides a way of extending standard reasons-responsive accounts of responsibility in a way that accounts for responsibility in cases of
unwitting omission. Additionally, I argued that there are good reasons to think that people can deserve blame in cases of unwitting omission and that cases of moral ignorance can also be considered failures of vigilance. This responds to some standard objections to reasons-responsive explanations of responsibility in cases of unwitting omission.
5.1 Introduction

You may not remember the last time you used the term ‘mind wandering’, but chances are you are familiar with the phenomenon. Consider three examples of mind wandering.

Oscar is a student taking a class on linear algebra. On Fridays, Oscar must attend a weekly review session with the class TA to go over the lecture material from the week. Even though Oscar really needs to focus (because he wants to do well on his upcoming exam), he periodically catches himself thinking about what he’s going to do over the weekend. As a result, he doesn’t fully catch some crucial review material.

Randy is driving home from work. He promised his wife that he would stop at the grocery store to pick some things up on the way home. While driving, his mind bounces between different topics: his brother’s funeral from the previous year, whether his kids need back-to-school clothes, and what to make for dinner next week. Because of this, Randy doesn’t notice that the car in front of him is stopping suddenly, causing Randy to get into a mild fender bender.

While these stories are fictitious, they do mimic results found in mind wandering research. There are well-documented performance deficits stemming from mind wandering associated with driving accidents (Yanko & Spalek, 2014), lecture
comprehension (Wammes et al., 2019), and simple (oftentimes occupational) tasks (Warm, Parasuraman, & Matthews, 2008).

These stories share a few things in common. First, each individual is engaged in a simple task. Second, the agent has some reason to care about performing the task well. Third, the agent’s mind drifts to something unrelated to the task that subsequently causes task performance to decline. Finally, the consequences of mind wandering are not good. These features map onto early predictions made about the conditions under which mind wandering is likely to occur. In a landmark article, Smallwood & Schooler (2006) cite three such conditions: (1) mind wandering is more likely to occur when an individual is engaged in a simple, monotonous task; (2) episodes of mind wandering are associated with diminished processing of external information because attention is internally focused, and; (3) mind wandering occurs without intention (see Smallwood & Schooler, 2006: 947).

In light of this proposal, early research on mind wandering converged on the idea that mind wandering is a form of task-unrelated thought (Smallwood & Schooler, 2015; Smallwood et al., 2003; Giambra, 1995; Schooler, 2002). In their original proposal, Smallwood and Schooler describe mind wandering as “…a shift of attention away from a primary task toward internal information,” and that the distinctive phenomenal characteristic of mind wandering is the “…experience [of] our minds drifting away from a task toward unrelated inner thoughts, fantasies, feelings, and other musings” (Smallwood & Schooler, 2006: 946). In a review paper written nearly ten years later, Smallwood & Schooler offer a similar characterization of mind wandering:

When the mind wanders, attention drifts from its current train of thought (often an external task) to mental content generated by the individual.
rather than cued by the environment. Often the thoughts that occur during mind wandering experiences are described as task unrelated or stimulus independent, terms that capture the independence of the experiences from perception and ongoing actions. Other terms, such as autobiographical thought or mind pops, capture the generative process that provides the content of the experiences itself. One term that captures the generative aspects of these experiences as well as their independence from perception is self-generated thought (2015: 489).

Most of the literature on mind wandering in cognitive science echoes this characterization of mind wandering as task-unrelated thought (hereafter, the TUT view). Recently, Mills and colleagues (2018) found that 96% of all papers on mind wandering published in 2016 defined mind wandering as task-unrelated thought. The TUT view also makes sense of the negative consequences of mind wandering. If mind wandering is characterized by involuntary shifts of attention away from an intended task, then the likely consequences of failing to attend to the task will be bad.

While the TUT view is the standard theory of mind wandering in cognitive psychology, there have been several issues raised against the TUT view. Four of these issues are notable. First, some claim that the internal dynamics of self-generated thought are distinctive of mind wandering, not task-relatedness (or the lack thereof) (Christoff et al., 2016; Irving, 2016; O’Callaghan et al., 2015). Irving, for example defines mind wandering as unguided attention, where attention drifts from topic to topic without monitoring or regulation (Irving, 2016: 563-64). Christoff et al. (2016: 719) propose that mind wandering is “a special case of spontaneous thought that tends to be more-deliberately constrained than dreaming, but less-deliberately constrained than creative thinking and goal-directed thought.” Call this the Dynamic view of mind wandering.
Second, some claim that mind wandering is a heterogeneous construct with multiple, overlapping attributes that no single instance of mind wandering fully exemplifies (Seli et al., 2018-c). The TUT view describes prototypical instances of mind wandering, according to Seli and colleagues, but other kinds of thought could be considered mind wandering to a greater or lesser degree depending on the level of resemblance between these cases of mind wandering and the prototypical case. Call this the Family Resemblances view.

Third, the TUT view apparently fails to categorize instances of mind wandering as mind wandering. For instance, it seems possible to mind wander in task-free scenarios. Someone lounging on the beach might find their thoughts drifting. This appears to be an instance of mind wandering. But, because the individual is not performing a task, the TUT view seems to preclude the possibility of mind wandering in this situation.

The fourth issue is that the TUT view seems unable to furnish a plausible solution to Irving’s Puzzle of the Purposeful Wanderer (Irving, 2016). The Puzzle arises from two facts about mind wandering: (a) mind wandering is a purposeless activity insofar as episodes of mind wandering are not driven by our goals, and; (b) mind wandering frequently advances our goals (Irving, 2016: 552). Given that TUT characterizes mind wandering as something that occurs in contradistinction to one’s goals and intentions, the TUT view cannot explain how mind wandering advances an individual’s goals.

These issues exploit overly simplistic formulations of the TUT view. This kind of misunderstanding was likely inevitable. To date, there has been no sustained elaboration or defense of the TUT view (such, perhaps, is the luxury and curse of being the standard view in a field). In this chapter, I aim to provide an account of the TUT view that
addresses these issues. The key to my account is the connection between mind wandering and vigilance, namely that mind wandering is a non-vigilant state. This provides material for a sophisticated account of the TUT view that also partly deflates the motivation for accepting alternative theories of mind wandering.

In Section 5.2, I offer a precise characterization of the TUT view. Section 5.3 outlines the possibility of mind wandering in task-free scenarios within the TUT framework. Section 5.4 discusses the Family Resemblances view of mind wandering and criticisms of the TUT view that advocates of the Family Resemblances view raise. Section 5.5 discusses the Dynamic view and criticisms of the TUT view that advocates of the Dynamic view raise. Finally, Section 5.6 outlines a solution to the Puzzle of the Purposeful Wanderer that is consistent with the TUT view. This also provides an occasion to connect my formulation of the TUT view back to the self-control as vigilance view outlined in Chapter 3.

5.2 The TUT View

The TUT view defines mind wandering in terms of three characteristics: task, task-unrelatedness, and thought. Each of these elements merits further discussion.

5.2.1 Task

Before we can understand what makes something task-unrelated, we need to understand what a task is. In particular, we need to understand what kind of tasks are such that individuals are susceptible to mind wandering when performing those tasks (hereafter, I will refer to this kind of task as a mind wandering-apt task). Cognitive scientists typically do not define the term ‘task’ (although see Wu, 2014: 83-84). Let’s
work with an example. Suppose there’s a graduate student relaxing on the beach listening to some music. While doing this, the graduate student decides to pass the time by moving a pile of small stones into a hole in the sand. In this example, is the graduate student performing four tasks (pursuing a degree, relaxing on the beach, listening to music, and moving stones), one task, or something in between?

Clearly, there are ways of carving the ontology of action where the graduate student is performing any number of tasks. The question here is, what view of tasks does the TUT view implicitly assume are mind wandering-apt? Not every agential activity is mind wandering-apt. In working through the TUT view of tasks, however, we can learn something important about the task properties that correlate with the propensity to mind wander.

Start with the activity of pursuing a degree. This is not a mind wandering-apt task. For one, pursuing a degree is a complicated task where pursuing the task goal is scattered over a long period of time. The task complexity leads to another point, which is that agents never directly pursue a degree. Instead, they pursue a degree by engaging in other activities (taking classes, writing papers, etc.). This turns out to be the key reason why pursuing a degree is not mind wandering-apt. The activity must be such that it can be pursued directly in order for it to be the kind of task where an agent can mind wander.

Next, consider lounging on the beach. We might exclude this as a mind wandering-apt task from the outset, but there are principled reasons for doing this. With relaxing on the beach, there’s no real distinction between the activity and the goal. To engage in the activity is to realize the goal (of relaxing), and one realizes the goal for as long as one engages in the activity. The lack of distinction between activity and goal,
then, precludes relaxing on the beach as a mind wandering-apt task. For some task to be mind wandering-apt, the task must be such that there’s a real distinction between realizing the goal of the task and undertaking certain means to pursue realizing the goal. Because there is a distinction between means and ends, this suggests that mind wandering-apt tasks are such that intending to perform the task entails adopting certain attentional commitments to focus on self-regulating one’s behavior in pursuit of the task goal. These attentional commitments are needed because pursuing the goal unfolds over time.

The temporal dimension of mind wandering-apt tasks leads to the final activity from the example, listening to music. Listening to music is not mind wandering-apt because realizing the goal does not take much time. We need to be more precise here. The precise task is *turn on some music* for purposes of listening to music. And turning on some music takes no time at all. Further, once the goal is realized, sustaining the goal does not require any further attention. You can turn on music and listen without focusing on any electronic equipment. This suggests that it must take some amount of time to move from means to end, or that sustaining the goal requires further attention from the individual. Of course, the amount of time is vague. The activity probably needs to take longer than a second, but probably doesn’t need to take longer than 7-10 minutes.³⁴

That leaves moving the stones as the only activity that is mind wandering-apt. It has all of the elements that make an individual performing the activity prone to mind wander. The goal of the activity is pursued directly, the means and ends of the activity are

³⁴ That invites an interesting empirical question: what is the shortest possible task where participants reliably report mind wandering during the task? To my knowledge, nobody has done any work on this question.
distinct, and the activity takes time to complete. This, in turn, provides some criteria for picking out mind wandering-apt tasks. Admittedly, this rules out a few tasks as being mind wandering-apt. But the resulting picture makes some intuitive sense. The activities we do where we are likely to mind wander are those tasks that take time, require attention, and are pursued directly. In other words, they are tasks that demand vigilance. They require monitoring, implementing a particular task schema, and regulating one’s behavior over time as task demands change.

This discussion reveals a connection between mind wandering and vigilance. Mind wandering is a non-vigilant state. To see this, consider again that mind wandering-apt tasks are those some agent pursues directly. Thus, in order to be susceptible to mind wandering, the agent must have an intention to perform some task from which the agent’s mind might wander. Specifically, the agent must have a proximal intention to perform the task, where a proximal intention is a mental state the having of which contributes immediately to its own execution (see Mele, Forthcoming). Put simply, a proximal intention concerns what an agent aims to do right now (the contrast is a distal intention, which is about what to do in the future; see Mele, 1992: 158). This gives us a different way to discuss excluding pursuing a degree from the category of mind wandering-apt tasks. Pursuing a degree is not something anyone can proximally intend to do.

Acquiring some proximal intention brings about commitments to guide attention toward considerations or task features relevant to performing the intended task. Reliably successful task performance requires maintaining these attentional commitments over time. This maintenance, as discussed in Chapter 2, is one aspect of the function of vigilance. Failure to maintain these commitments, for instance, constitutes a failure of
vigilance (assuming that the individual is unimpeded and has neither revised nor abandoned her intention).

This makes explicit the connection between mind wandering and vigilance. Mind wandering-apt tasks are those that require vigilance for reliably successful performance. Mind wandering is a failure of vigilance insofar as it involves failure to maintain attentional commitments necessary for reliably successful task performance (these failures of vigilance might be locally useful, an issue covered in Section 5.6).

This also provides a preliminary indication of the thoughts relevant to mind wandering. Task performance requires some kind of action-guiding cognitive state (Shepherd, 2014). These are relevant to mind wandering, but may not be constitutive of mind wandering. That is because we want to identify states that are unrelated to task-relevant thoughts, not just what task-relevant thoughts are. So we need an account of task-relatedness and task-unrelatedness before we can characterize the relevant class of thoughts.

5.2.2 Task-Relatedness

What makes a thought task-unrelated? As a first pass, we could propose that any thought irrelevant to task performance is unrelated. But this pushes the question back to what makes a thought irrelevant to task performance. So we need to aim for something more informative.

Here’s a more serious proposal:

(U1) A thought that T is unrelated to task A for S when S’s thinking that T precludes her from having some thought that R, where S’s thinking that R is necessary for S to A.
This is too strong. One can mind wander while driving without crashing. In other tasks, as well, mind wandering does not cause task performance to fall to 0. So task-unrelatedness cannot be a matter of precluding thoughts necessary for task performance.

One fix concerns the relevant kind of task performance. Someone might successfully complete a task accidentally, even while mind wandering. Suppose, for instance, that someone has the task of arranging lettered blocks into patterns that spell out words. This task might be accomplished accidentally while mind wandering. But, mind wandering will decrease the likelihood of successful performance on a given task, especially when the task is iterated. Thus, we should amend (U1) to include a phrase about reliable performance:

(U2) A thought that T is unrelated to task A for S when S’s thinking that T precludes her from having some thought that R, where S’s thinking that R is necessary for S to A reliably.

This, however, does not fully address the concern. Given the high frequency of mind wandering in daily life (Killingsworth & Gilbert, 2010), it’s safe to say that people can mind wander and perform some tasks reliably.

Let’s start from a different place. Consider the following account of what makes a thought task-related:

(R1) A thought that T is related to S’s A-ing when S’s thinking that T contributes to her A-ing.

With (R1) on the table, we can pose the question of what makes it the case that a thought contributes to some activity. One natural suggestion, building on the need to secure reliable performance, is that a thought contributes to someone’s activity when the individual’s having that thought increases the reliability of successful task performance.
Many tasks can be routinized and off-loaded to habits. However, given dynamic and unpredictable changes to the environment, routinized task performance requires monitoring for changes that signal the need to shift parameters of performance. Monitoring, then, enables flexible task performance. This monitoring is another aspect of the function of vigilance. Hence, we’ve uncovered another connection between mind wandering and vigilance. Agents exhibit vigilance when they are disposed to acquire cognitive states that contribute to goal-directed behavior. Mind wandering, as a manifestation of non-vigilance, decreases an agent’s monitoring and, hence, decreases the likelihood that the agent acquires cognitive states that contribute to goal-directed behavior.

This also clarifies what we need from an account of task-unrelatedness. Given that the task-relatedness of a thought is a function of whether that thought is such that a vigilant agent would, when appropriate, acquire that thought, a task-unrelated thought is one that precludes monitoring, not one that precludes an individual from having particular thoughts.

Specifying this relation is difficult, in part because much remains unknown about monitoring. However, the outlines of an empirical research program can be derived from the discussion of control in Chapter 3. If we assume the correctness of the shared resource theory of control, then we know that cognitive processes that draw on similar representational resources cannot be engaged simultaneously. Further, when a process consumes a particular representation (or members from a fixed representational class), there are costs associated with switching to a different cognitive process that relies on the same representation (or some member of the same class). Unless monitoring is actively
maintained, detection of cues relevant to task switching (when the tasks draw from the same representational pool) is less reliable. So, we can get a grip on the precise nature of task-unrelatedness by understanding the basic representational code used in cognition and the allocation of those resources over time. This sort of computational program is currently in its infancy (however, see Eliasmith et al., 2012 and Gershman, Horvitz, Tenenbaum, 2015), but over time we should discern a more concrete understanding of what makes certain states task-unrelated given an individual’s performance of some task.

The connection between monitoring and mind wandering indicates the needed changes to (U2). First, we can define ‘contribution’ in terms of vigilance:

(C) A thought that T contributes to S’s A-ing when S’s thinking that T manifests vigilance with respect to her A-ing.

If manifesting vigilance with respect to some activity amounts to reliable performance of that activity, then we can amend (R1) along the following lines:

(R2) A thought that T is related to S’s A-ing when S’s thinking that T is part of S’s A-ing reliably.

From this, we can amend (U2):

(U3) A thought that T is unrelated to S’s A-ing when S’s thinking that T manifests a lack of vigilance with respect to her A-ing.

Or, more precisely:

(U3) A thought that T is unrelated to S’s A-ing when S’s thinking that T precludes her from monitoring for consideration C, where C is such that, were C detected, it would result in S’s having some thought R, where R contributes to S’s A-ing reliably.

This proposal of task-unrelatedness has a number of virtues. First, an account of the thoughts relevant to mind wandering falls out of the proposal. The thoughts constitutive of TUT are those that preclude monitoring. Second, (U3) indexes task-unrelatedness to
individual performances rather than thinking of task-unrelatedness as holding between thoughts and tasks. What precludes one individual from monitoring might not preclude another individual from monitoring. This fits well with research on trait-level variables associated with the propensity to mind wander (Seli, Risko, Smilek, 2016) and the relationship between working memory capacity and mind wandering (McVay & Kane, 2010). Third, (U3) meshes with current methodology used to study mind wandering. In particular, (U3) makes sense of the fact that psychologists use iterated tasks in mind wandering experiments in order to test reliable performance, the fact that these tasks unpredictably shift conditions for successful performance throughout the experiment (thereby testing monitoring), and the fact that most cognitive scientists operationalize mind wandering in accordance with the TUT view.

The TUT view also implies that mind wandering is a form of non-vigilant thought. As explained above, this means that an individual is mind wandering only if the individual has some task-unrelated thought. Is this also a sufficient condition for mind wandering? Not quite. Sometimes, task-unrelated thought results from attentional capture, and these instances do not seem to be cases of mind wandering.

Attentional capture is a phenomenon where some stimulus forces a reaction that shifts focus toward the suspected source of the stimulus. For example, if you’re sitting in a café or restaurant and hear a loud bang, your attention is immediately drawn to the suspected source of the noise. This is automatic in the same way that your leg automatically kicks out when the doctor hits your knee with a reflex hammer.

Instances of attentional capture are not mind wandering despite the fact that they do result in task-unrelated thought. Thus, task-unrelated thought is not sufficient for mind
wandering. To get that, we need to add a stimulus-independence condition. What is distinctive about attentional capture is that stimuli bring about instances of attentional capture. Mind wandering, on the other hand, is a kind of self-generated thought. So, adding a stimulus-independence condition rules out attentional capture from counting as mind wandering. On this modified TUT view, then, an individual is mind wandering if and only if the individual has a stimulus independent task-unrelated thought (cf. Stawarczyk et al., 2013).

The stimulus independence condition deserves some clarification. Oftentimes, experience contains a mixture of stimulus-dependent and –independent thought. It is not the case that mind wandering bears no connection to stimulus-dependent thought; rather, it’s that stimulus-dependent thoughts are not themselves constitutive of mind wandering. For example, suppose I’m listening to a lecture and I see someone eating an orange. That makes me think of Florida, which makes me think of my late grandfather and the time I drove him from Lancaster, Pennsylvania to Boynton Beach, Florida (true story, by the way). In this case, a stimulus-dependent thought results in my thoughts drifting, but the mind wandering does not begin until I think of Florida. While we can draw a theoretical distinction between stimulus-dependent and stimulus-independent thought, however, in practice it is likely the case that mind wandering results from episodes of stimulus-dependent thought.

This concludes the statement of the TUT view. We now have criteria for identifying mind wandering-apt tasks, the relation of task-relatedness (and, likewise, the relation of task-unrelatedness that is relevant to mind wandering), and the kinds of thoughts constitutive of mind wandering episodes. Mind wandering-apt tasks are those
that require vigilance for reliable performance. Following on this, what makes certain
thoughts task-unrelated (and, hence, partly constitutive of mind wandering) is that one’s
having such thoughts precludes one from being vigilant in ways that support task
performance. This is the core element of this formulation of the TUT view.

5.3 Task Free Scenarios and Mind Wandering

Mind wandering can occur in task free scenarios. Consider someone lounging on
the beach letting her thoughts drift. This might be an instance of mind wandering. The
TUT view entails that if an individual is mind wandering, then that individual has at least
one stimulus-independent task-unrelated thought. However, the beach lounger is not
performing a task (unless we construe tasks so liberally so as to include relaxing on the
beach as a task). So the TUT view rules out the possibility that the beach lounger can be
mind wandering (Irving, 2016: 554). This seems to be a counterintuitive upshot of the
TUT view.

The mistake in the objection, however, is to assume that the TUT view rules out
the possibility of mind wandering because there is no task being performed. This
misconstrues the TUT view as entailing that if an individual is mind wandering, then the
individual is performing a task and the individual has at least one task-unrelated thought.
As we saw above, the TUT view is best understood as claiming that mind wandering is a
non-vigilant state. Certain patterns of thought preclude monitoring, and these are partially
constitutive of mind wandering episodes. This lack of monitoring entails a diminished
propensity to acquire action-relevant cognitive states. An individual could have this
diminished propensity even in task free scenarios. For example, the beach lounger might
have certain standing intentions, where these are dispositions to acquire certain intentions in particular circumstances (see Mele, 2007: 736-38). When these circumstances obtain, the beach lounger (while mind wandering) is less likely to detect this and acquire the relevant intentions. This is one way in which an individual in a task-free scenario might be non-vigilant.

This elaboration of the TUT view, when applied to mind wandering in task free scenarios, explains the way in which mind wandering is a form of disengaged cognition. Task-unrelated thought is characterized by the agent’s diminished reactivity to dynamic environments. This is important because it implies that task-unrelated thought is not a matter of being in a particular mental state; rather, task-unrelated thought is a kind of inflexibility or difficulty in switching focus and acquiring cognitive states related to alternative task sets.

The precise amendment to the TUT view to account for the beach lounger case is tricky to pin down. Roughly, the idea is that one could have a policy to form proximal intentions to A when in C. So, an individual might be in a state that greatly lowers the likelihood of detecting that she is in C, and thereby unwittingly fail to acquire an intention to A.\(^{35}\) This failure counts as a form of task-unrelated thought because we can construe policies as generalized task kinds. Hence, this might be the kind of task-unrelated thought is characterized by the agent’s diminished reactivity to dynamic environments. This is important because it implies that task-unrelated thought is not a matter of being in a particular mental state; rather, task-unrelated thought is a kind of inflexibility or difficulty in switching focus and acquiring cognitive states related to alternative task sets.

\(^{35}\) I say the formulation is tricky to pin down for this reason. There are many factors that alter the likelihood of some individual noticing or being aware of some change in the environment. In order to determine which factors are relevant, one would need to set a baseline probability for detecting changes that determines which factors cause a downward shift in probability. Then, there would need to be some principle that excludes external factors. The rough idea, however, seems intuitive enough. There are standards of successful awareness that set our expectations for what people notice about their surroundings. Intrinsic properties of the agent that negatively affect the probability of someone living up to those standards are candidates for constituting mind wandering episodes.
unrelated thought that the beach lounger exhibits. Even though she is not currently performing some focal task, she still fails to exhibit vigilance in a way and, partly in virtue of that, counts as mind wandering.

5.4 Intentional Mind Wandering and the Family Resemblances View

Recently, Paul Seli and colleagues (2018-c) have argued that every proposed definition of mind wandering is extensionally inadequate. They consider the 5 most commonly proposed definitions of mind wandering (task-unrelated thought, unintentional thought, stimulus-independent thought, stimulus-independent task-unrelated thought, and meandering/unguided thought) and show that they are either too inclusive or too exclusive to be adequate definitions.

The core of Seli and colleagues’ argument is that every proposed definition of mind wandering is inadequate. When discussing the TUT view, Seli and colleagues list three inadequacies: (1) the TUT view precludes task-free mind wandering; (2) some stimulus-associated thought might constitute an episode of mind wandering, which is inconsistent with the version of the TUT view presented in Section 5.2, and; (3) the TUT view rules out the possibility of intentional mind wandering. These considerations form part of the argument for the Family Resemblances view. Roughly, the TUT view adequately characterizes certain kinds of mind wandering but not others, namely intentional and stimulus-associated mind wandering. Since no single definition encompasses all kinds of mind wandering, there must be a relation of family resemblance that obtains between different characterizations of mind wandering that jointly constitute the mind wandering category.
The first two considerations, however, rest on simplistic versions of the TUT view. Section 5.3 outlines how the TUT view explains the possibility of mind wandering in task-free scenarios.

Seli and colleagues argue that the TUT view is too exclusive because the view excludes stimulus-associated thought from counting as mind wandering. They give the following example:

Consider, for example, the scenario in which you are eating dinner with your family and your thoughts drift to an argument you just had with your sibling, who is seated next to you. Such stimulus-associated thoughts would qualify as ‘mind-wandering’ to many people. Yet, because these thoughts were triggered by (and continue to feature) an environmental stimulus, the [TUT view] rejects them as mind wandering (2018: 481).

As we saw in Section 5.3, however, stimulus association does not rule out the possibility of mind wandering on the TUT view. Instead, stimulus-dependent thoughts that lead to stimulus-independent thoughts are candidates for constituting episodes of mind wandering. So the TUT view, when properly characterized, turns out to be extensionally adequate in this respect.

The issue of intentional mind wandering, however, is more serious. Consequently, the possibility of intentional mind wandering is the biggest consideration in favor of the Family Resemblances view. This is because if the TUT view is correct, then there can be no intentional mind wandering. Hence, if there is such a thing as intentional mind wandering, then the TUT view is false. Let’s consider first the evidence in favor of intentional mind wandering before considering whether intentional mind wandering threatens the TUT view.

Some cognitive scientists have recently argued for the importance of distinguishing between unintentional and intentional mind wandering, where an agent can
either involuntarily shift attention to task-unrelated thoughts (unintentional mind wandering) or, instead, engage higher-order, deliberate control and choose to shift attention (intentional mind wandering) (e.g., Seli, Risko, & Smilek, 2016; Seli, Risko, Smilek, & Schacter, 2016; Golchert et al., 2017). The basis of the distinction is that these two kinds of mind wandering independently predict variables of interest (Seli, Risko, & Smilek, 2016).

Some researchers have found different associations of trait- and state-level variables between unintentional and intentional mind wandering. On this basis, these researchers argue for the theoretical usefulness of distinguishing intentional from unintentional mind wandering (see Robison & Unsworth, 2018; Seli, Carriere, & Smilek, 2015; Vannucci & Chiorri, 2018). For instance, Seli, Risko, & Smilek (2016) asked participants to complete either a difficult or easy version of the Sustained Attention to Response Task, a task that requires participants to respond to frequently presented items and withhold a response for infrequent target items. Mind wandering was measured with randomly distributed thought probes that prompted participants to report whether they were unintentionally or intentionally mind wandering at a given moment. Results showed that when collapsing across the two types of mind wandering, there was no significant difference in the overall rates of mind wandering between difficult and easy conditions. However, when analyzing intentional and unintentional mind wandering separately, the authors found: (1) higher rates of reported unintentional mind wandering than intentional mind wandering; (2) higher rates of reported intentional mind wandering in the easy-SART condition relative to the difficult-SART condition, and; (3) significantly higher proportion of reported unintentional mind wandering in the difficult-SART condition.
relative to the easy-SART condition. One significant conclusion that Seli and colleagues draw from this is: “...had we ignored the distinction between intentional and unintentional mind wandering, we would have drawn the incorrect conclusion that mind wandering remains unchanged across conditions” (2016: 688).

Additionally, participants often self-report intentional mind wandering when prompted to distinguish between task-related thoughts, unintentional, and intentional mind wandering (e.g., Seli, Carriere, & Smilek, 2015; Seli, Risko, & Smilek, 2016; Seli et al., 2017; Seli et al., 2018-b). The fact that people self-report intentional mind wandering is _prima facie_ evidence in favor of there being such a thing as intentional mind wandering.

The TUT view is incompatible with the possibility of intentional mind wandering. There is no way to intend to have task-unrelated thoughts, as any thoughts that one focuses on intentionally will constitute task-related thoughts. Because mind wandering consists in a diminished propensity to acquire task-relevant thoughts, any thoughts that one has in virtue of intending to have such thoughts will necessarily not be connected to a diminished propensity to acquire task-relevant thoughts. Thus, the thoughts cannot be unrelated and so cannot constitute mind wandering. On the TUT view, intentional mind wandering is impossible (see Murray & Krasich, Manuscript) for a sustained argument in defense of this point).

On its face, the notion of intentional mind wandering is paradoxical. Mind wandering seems to be inherently passive, as Zac Irving notes:

> It seems essential that mind-wandering lacks purpose; almost by definition, it contrasts with goal-directed forms of cognition like planning
a trip or solving a crossword. Consider the term ‘mind-wandering’ itself. Wandering is a purposeless movement…To say that someone’s mind is wandering, then, implies that her thinking is purposeless; it is not developing toward a goal or end-point (2016: 549-50; see also Dorsch, 2015).

Mind wandering also occurs without much effort (Sripada, 2016), attention (Metzinger, 2013), or working memory demands (McVay & Kane, 2010). Intentional activity, on the other hand, is goal-directed and thereby paradigmatically active.

Recent work on the neuroscience of mind wandering and goal-directed activity reinforces this distinction. One meta-analysis showed that mind wandering, episodic future thinking (envisioning forthcoming events), and personal goal processing, despite sharing some neural circuitry, have distinct neural correlates (Stawarczyk & D’Argembeau, 2015). Specifically, mind wandering was associated with diminished dorsolateral prefrontal cortex—a brain region often used to measure levels of cognitive control—recruitment relative to episodic future thinking and personal goal processing. This suggests differential recruitment of attentional control abilities for these activities. Stawarczyk and D’Argembeau propose that episodic future thinking and personal goal processing can be components of mind wandering, but that these forms of internal thought are inherently distinct. Although this meta-analysis does not distinguish between unintentional and intentional mind wandering, it reinforces the claim that intentional mind wandering is paradoxical.

As noted earlier, it seems theoretically useful to distinguish intentional and unintentional mind wandering given that they differentially predict various state- and

---

36 This does not imply that goal-directed activity requires effort. However, typical instances of goal-directed activity require more effort than typical instances of mind wandering.
trait-level variables. This data is consistent with a weaker conclusion, namely that in studies of mind wandering researchers must distinguish mind wandering from other, intentional mental activities that share certain superficial properties of mind wandering. On this interpretation, the study from Seli and colleagues show that participants performing certain lab-based tasks are disposed both to mind wander and to engage in other forms of distracted mental activity. Researchers should be careful not to lump mind wandering and these other deliberate forms of distracted cognition together in their analyses. The decision to label this deliberate form of distraction ‘intentional mind wandering’ is a theoretical choice that Seli and colleagues are imposing on the data.

There is some motivation to adopt this weaker interpretation. This would require, among other things, construing instances of intentional mind wandering as instances where an agent chooses to disengage from an external, ongoing task and engage in another—already well-defined—mental activity, such as personal goal processing, episodic future thinking, motivated task switching, focused daydreaming, and others.

Consider, as an example, the similarities between intentional mind wandering and focused daydreaming. As Dorsch (2015: 792) describes it, focused daydreams are cases where we “withdraw from the world surrounding us and start to experience or think about objects and events that are largely absent, past or imaginary. But we take a much more active part in directing the order and content of our mental episodes and usually stick to a particular topic or issue.” And this seems to map onto descriptions of what it’s like to intentionally mind wander. Thus, purported cases of intentional mind wandering might just be cases of focused daydreaming. This assumes that intentional mind wandering and focused daydreaming are distinct, and there are good reasons to hold this assumption.
Dorsch (2015: 807) notes that the content of focused daydreams are narratively structured in a way that mind wandering content is not. This is consistent with Seli et al.’s (2017: 152) findings that intentional mind wandering correlates with significantly lower reports of vagueness (as compared to unintentional mind wandering). This also reflects a deeper difference, which is that focused daydreaming involves exercises of imaginative mental agency, whereas mind wandering is inherently non-agentive (cf. Dorsch 2015: 792). This is one reason to resist assimilating the categories of daydreaming and mind wandering, and hence a reason not to equate focused daydreaming and intentional mind wandering.

This is not to say that every purported instance of intentional mind wandering is just a case of focused daydreaming. Consider, for example, a case where somebody is taking a road trip and currently driving down a long, flat stretch of highway in central Idaho. This individual might stop consciously attending to the road and begin thinking about a paper she’s writing or a talk that she’s giving in a few weeks. While there are some similarities, this doesn’t seem like daydreaming. This might seem to be a case of intentionally mind wandering, but an alternative description is available.

Repeated performance of some activity makes that activity overlearned. When some activity is overlearned, executing the activity as a task actually hinders performance. Here, a task is an activity performed under the guidance of an agent’s executive resources, like attention and cognitive control (see Mole, 2011: 51).

Christopher Mole (2011: 62) explains:

It is typically harder to give sustained attention to familiar, well-understood tasks than it is to give attention to tasks that are poorly understood. The thought...is that giving one’s attention to a well-understood task involves marshalling a large set of resources, just because the task is so well-understood.
When there are too many available resources, executing an activity as a task (where this implies guided, attentive performance) hurts performance. Consider, for instance, that tying one’s shoes or walking up stairs is more difficult when one focuses on every component of the activity. Thus, there is some incentive to switch tasks when one is performing an overlearned activity. This might be what’s happening in the driving case. Driving is an overlearned activity, and the roads in central Idaho do not demand much attention. So there are good reasons to switch focus to some other task that does benefit from sustained attention (more on this in Section 5.5). On this view, the distracted drive is not intentionally mind wandering, but instead engages in motivated task switching. This is not to say that every instance of intentional mind wandering is either an instance of focused daydreaming or motivated task switching; rather, the examples above indicate how one might go about reducing purported cases of intentional mind wandering to other well-defined categories of mental activity.

These suggestions, however, conflict with the fact that people consistently self-report intentionally mind wandering. Why, if there is no such thing as intentional mind wandering, do studies of intentional mind wandering elicit such self-reports? We might be tempted to chalk this up to errors in folk thinking. People are bad at introspecting (or categorizing mental phenomena on the basis of introspection), and so we should dismiss self-reports as evidence for theory construction.

We can, however, deliver something more concrete in this case. One possibility is that participants have a mistaken conception of mind wandering based on experimental instructions. In an experimental context, participants are told to perform the task provided by the experimenter. Thus, when participants have any thought unrelated to the
experimental paradigm, they categorize this thought as mind wandering. Participants then might report intentional mind wandering because they chose to think about something unrelated to the experimental task, as per the experimental definition of intentional mind wandering. Notice, however, that in this example the participants’ understanding of what a task is diverges from the notion of task used in unpacking the TUT view. When we use this notion of task, these thoughts are not mind wandering.

This might seem like a natural place to object. Typically, tasks (at least in research on mind wandering) are restricted to researcher-imposed tasks completed in a laboratory. We can make sense of intentional mind wandering with this conception of task, because intentional mind wandering is just intentionally thinking about something unrelated to the researcher-imposed task. The participants can acquire a new task through this intention, but she also retains the researcher-imposed task. This, however, threatens to make the mind wandering construct too expansive. If mind wandering occurs whenever one fails to think of something they should think about, then any distracted state of mind would qualify as a mind wandering state. For example, procrastination and mind wandering seem to be distinct mental phenomena, but under this conception of task the two are coextensive. What characterizes mind wandering is having some thought that fails to contribute to executing an intended task. Hence, having an intention is essential to being able to mind wander. In laboratory settings, researchers give study participants reasons to adopt certain intentions, but these participants are able to intentionally swap out those researcher-incentivized intentions for other intentions.

The reason for focusing on the existence of intentional mind wandering is that the possibility of intentional mind wandering is a prominent feature of the argument for the
Family Resemblances view. This is because the possibility of intentional mind wandering is inconsistent with the TUT view.

One objection is that the responses to the Family Resemblances view on behalf of the TUT view all presume the truth of the TUT view. But this seems to beg the question against the Family Resemblances view. Why should we opt to favor the TUT view in the first place? One response is that the TUT view is the standard view and, absent any strong considerations in favor of revising the view, there is no reason to consider alternative definitions as on a par with the TUT view. Another response is that the TUT view makes fewer controversial assumptions about psychological theorizing and methodology. Briefly, Seli and colleagues rely on the idea that the term ‘mind wandering’ refers to a human-created category, similar to terms like ‘game’ and ‘chair’ (see Seli et al., 2018-c: 482-83). These categories tend to have fuzzy boundaries, so we should expect the same to be true of mind wandering. This, however, is incorrect. Psychological phenomena are part of the fabric of reality, and the associated constructs employed in psychological theory aim to map onto these phenomena. For this reason, we should want (and expect) our psychological categories to have their membership conditions spelled out in terms of necessary and sufficient conditions. This reflects the fact that psychological categories—unlike, say, economic categories—are discovered, not created. Further, there is a way to regiment our categorization of various cognitive phenomena under discussion, outlined above, that does not require the concessions that Seli and colleagues make in their Family Resemblance proposal. For all of these reasons, the TUT view seems preferable to the Family Resemblances view.
5.5 Empirical Credibility and the Dynamic View

Another recent trend in mind wandering research is to focus on the dynamics of spontaneous thought as the basis for defining mind wandering. Christoff et al. (2016), Sripada (2016-b; Forthcoming), and Irving (2016) have developed this kind of view. Recall that the Dynamic view characterizes mind wandering in terms of the dynamics of the train of thought constitutive of an episode of mind wandering. Mind wandering, on this view, is a kind of moderately constrained spontaneous thought or unguided attention.

One problem with the Dynamic view is that the position risks making mind wandering into an empirically intractable phenomenon. The TUT view reflects a construct of mind wandering that is amenable to study in laboratory conditions. The Dynamic view utilizes certain modal notions (like moderate deliberative constraint) that are impossible to identify in laboratory conditions. If it is not possible to determine empirically whether the necessary conditions on mind wandering obtain, then mind wandering ceases to be an object of scientific study. Thus, the Dynamic view appears to violate a constraint on empirical credibility. We do not want the construct of mind wandering to be irreducibly subjective, because then the construct is inaccessible to standard experimental methods. Hence, the construct of mind wandering should avoid inclusion of empirically inaccessible phenomenon.

Noting this empirical credibility constraint reveals that the Dynamic view is subject to a dilemma (see Metzinger, Forthcoming for an alternative presentation of a similar dilemma). On the dynamic theory, one of the properties characteristic of mind wandering episodes is a suitable degree of constraint on the sequence of thoughts constitutive of mind wandering. As Christoff et al. (2016: 719) note, mind wandering is
“a special case of spontaneous thought that tends to be more-deliberately constrained than dreaming, but less-deliberately constrained than creative thinking and goal-directed thought.” Constraint, as used to determine the relevant dynamics of thought, is either a subjective or objective property (where a subjective property is such that the conditions under which one has the property are accessible only to introspection). If constraint is an objective property, then whether some individual exhibits the property is determined externally in ways that are introspectively inaccessible to the individual that has the property. If the occurrence of mind wandering can only be determined externally, then mind wandering episodes are no longer introspectively accessible (in other words, whether or not you are mind wandering is not something you can know for yourself on the basis of introspection). That’s a counter-intuitive claim, and it also jeopardizes the dominant methodologies of current empirical research programs, because most mind wandering research utilizes thought probes (Smallwood & Schooler, 2015; Weinstein, 2018). However, if constraint is a subjective property, then the dynamic theory seems to jeopardize the empirical study of mind wandering because mind wandering would be irreducibly phenomenological. This is problematic because it implies that mind wandering is not so much a state of the brain, but instead a way that one experiences certain conscious episodes (or the transitions between such episodes). The problem here is that this rules out the possibility of being mistaken about whether or not one is mind wandering. It seems, however, *prima facie* plausible that one might confuse mind wandering with some other cognitive state, which conflicts with the idea of mind wandering as an essentially subjective construct (consider, as contrast, the impossibility of being mistaken about whether one is in pain).
Hence, on both horns of the dilemma, the dynamic theory has difficulty with the empirical credibility constraint. The TUT view, whatever its other shortcomings, has solid empirical credentials that its dynamic counterpart lacks.

Proponents of the Dynamic view raise several objections to the TUT view. In particular, Irving (2016) has offered a number of significant criticisms of the TUT view on behalf of his version of the Dynamic view. The first challenge questions whether some task-related thoughts might constitute episodes of mind wandering:

Wandering thoughts are sometimes related to the (occurrent) task from which one’s mind wanders. Suppose that my mind is wandering as I walk to campus. At some point, I recall that there’s a disguised rock in the grass through which I’m walking. Rather than stop my mind from wandering, this recollection changes the course of my wandering thoughts: I muse that it’s lucky I remembered the rock, recall the many times I’ve tripped over it, imagine myself sprawled face-first on the grass, etc. Fortified by my wandering thoughts, I successfully remember the rock when I reach it, and hop over it unscathed. Here, my wandering thoughts seem related to the occurrent task I perform as my mind wanders: namely, walking to campus. Thus, mind-wandering is sometimes related to one’s occurrent task (Irving, 2016: 555).

This is a version of the challenge raised by Seli et al. (2018-c) that some stimulus-dependent thought might constitute episodes of mind wandering. In Irving’s example, however, the direction of causality runs in the other direction. Stimulus-independent thoughts result in the individual having stimulus-dependent thoughts (about a particular rock on one’s route across campus) that then contribute to some later activity of the agent. But this is no problem for the TUT view as outlined in Section 5.2, as that view allows for the possibility of interactions between stimulus-dependent and stimulus-independent forms of thought.

The second challenge concerns the notion of task that the TUT view employs. Irving offers an example of Martha, who is balancing her stock portfolio when her
thoughts drift to thinking about a cookbook that she’s writing. Given that Martha is in the process of writing a cookbook, her cookbook-related thoughts do not appear to be task-unrelated in the way necessary for mind wandering. Hence, Irving concludes that the TUT view, on its own, does not count Martha’s drifting thoughts as mind wandering. The only way to do this is to invoke assumptions about the relationship between (lack of) guidance and mind wandering, but these assumptions are not available to the TUT view. However, this relies on an overly simplistic characterization of the TUT view. As we saw in Section 5.2, the TUT view is able to draw out substantial connections between mind wandering and vigilance. In particular, the TUT view implies that mind wandering is a non-vigilant state. And insofar as vigilance is related to the guidance of thought over time, the TUT view can appeal to notions of guidance without making substantive assumptions. So, this criticism of the TUT view fails against more sophisticated versions of the TUT view.

Finally, Irving notes that the TUT view cannot distinguish mind wandering from rumination and absorption, two cognitive phenomena that are distinct from mind wandering. As he uses the terms, rumination is a form of passive obsession with different kinds of distress and the consequences of that distress, whereas absorption involves heightened singular focus on one particular topic or consideration (Irving, 2016: 559-60). Irving thinks that the TUT view cannot distinguish mind wandering from rumination and absorption, in part because rumination and absorption are frequently unrelated to one’s current tasks (especially when these conditions manifest in clinical populations) (Irving, 2016: 560). However, once we see that mind wandering is a non-vigilant state (on the TUT view), this objection loses its force. Rumination, absorption, and mind wandering
are related in being non-vigilant states. However, rumination and absorption are non-vigilant in virtue of violating norms of efficiency in cognition (outlined in Chapter 1). Mind wandering is non-vigilant in terms of manifesting a lack of monitoring and maintenance in one’s behavior. Hence, we can derive a principled distinction between mind wandering, rumination, and absorption from various features of vigilance.

Someone might object that, when we invoke aspects of vigilance to explain mind wandering on the TUT view, there turns out to be little difference between the Dynamic view and the TUT view. This seems correct. Especially on Irving’s view, where mind wandering is unguided attention, the TUT view (as spelled out in connection to vigilance) implies that mind wandering is a form of unguided thought. This, I think, indicates what seems so intuitively correct about the Dynamic view. However, the TUT view can incorporate these aspects of the Dynamic view without appealing to empirically intractable notions (recall that various dimensions of vigilance are empirically tractable in ways spelled out in Chapter 3). Hence, rather than being a merely verbal dispute, this discussion shows that the TUT view does not differ much from the Dynamic view, and where it does differ, the TUT view fares better than its counterpart.

5.6 Mind Wandering and Control

In his (2016), Zac Irving outlined the Puzzle of the Purposeful Wanderer and set it as a challenge that any theory of mind wandering ought to resolve. The puzzle revolves around two claims: (1) mind wandering is a purposeless activity insofar as episodes of mind wandering are not driven by our goals, and; (2) mind wandering frequently advances our goals (Irving, 2016: 552). Evidence for (1) comes from the fact that mind
wandering is associated with deficits across a range of different activities and that paradigmatic instances of mind wandering tend to be unrelated to tasks one is currently performing. Evidence for (2), however, comes from recent research that reveals numerous beneficial consequences of mind wandering. Mind wandering, in some contexts, is associated with planning benefits (see Baird et al., 2011). That is, mind wandering frequently drifts to one’s future goals, and simulation of activities needed to achieve those goals results in better performance, and this simulation seems to occur during episodes of mind wandering (Sripada, 2016-b). Additionally, mind wandering is associated with higher levels of creativity (Baird et al., 2012) and benefits to goal-directed thinking (Gorgolewski et al., 2014). Hence, mind wandering does, at least in some contexts, appear to contribute to planning agency. This leaves us with two questions. First, how might mind wandering contribute to planning agency given that it is a non-vigilant state? Second, how does the TUT view resolve the Puzzle of the Purposeful Wanderer? Given that the TUT view easily explains the inherent purposelessness of mind wandering, this second question amounts to how the TUT view explains the fact that mind wandering frequently contributes to planning agency.

On the TUT view, mind wandering is neither an exercise of mental agency nor an intentional activity. However, mind wandering is connected to control, and this connection explains the relationship between values/goal-directed behavior and mind wandering. To see the connection, we need to return briefly to the view on the intrinsic costs of control outlined in Chapter 3. Consider a normal individual with a hierarchy of values, goals, and commitments to be pursued over time. Now suppose that there are various configurations available for that individual, where a configuration corresponds to
a particular allocation of cognitive control in service of goal-directed behavior. Each configuration is associated with an intrinsic cost, where that cost is a function of the value of one configuration weighed against other possible configurations. One possible configuration is a disengaged or ‘plural pursuit’ stage, where the individual does not focus on performing one task at the exclusion of others, but rather drifts between focusing on multiple goals. As we saw in Chapter 3, for some individual engaging in goal-directed behavior, the opportunity costs associated with that behavior shift over time. These shifts help to explain the transition to a state of mind wandering. When the opportunity cost of engaging in that behavior (i.e., maintaining a specific configuration) falls below a certain threshold, and the opportunity cost associated with plural pursuit rises above a certain threshold, the probability of plural pursuit increases. This state of plural pursuit corresponds to mind wandering insofar as plural pursuit is a relatively unstructured mode of cognition. That is, when opportunity costs change and begin to favor the ‘plural pursuit’ configuration, there is a higher likelihood that control is reallocated to favor a configuration conducive to mind wandering.

This connection to control provides a framework for explaining the second aspect of the Puzzle of the Purposeful Wanderer. Opportunity cost calculations are structured by one’s commitments, goals, and values. Hence, allocations of control are informed by commitments, goals, and values. In turn, whether one is mind wandering, and the content of that mind wandering, is partly determined by one’s values, goals, and commitments. While this provides a good start on explaining the value-directed benefits of mind wandering, there is still the issue of why individuals mind wander at all. Given the
benefits of structured, goal-directed, focused cognition, why would we develop the tendency to mind wander in the first place?

Return to the notion of overlearned activities discussed in Section 5.4. It is typically more difficult to perform an overlearned activity, and performance on these activities tends to decrease when one is overly attentive to performing the activity. So, in these cases, it is beneficial to disengage and switch focus to something else. However, rather than shifting focus to some other discrete, well-defined activity, it can be beneficial to engage in a kind of random search for future activities that one could think about for purposes of consolidation or simulation (see Sripada, Forthcoming). Consider an example of walking across campus from your car to your office. You don’t need to think about every step you take or which way to go because you’ve done it a million times before. So it’s useful to disengage and let that activity be governed by one’s habits and routines while focus shifts to other goals. This explains why, with respect to one’s walking, it is sometimes beneficial to mind wander while walking despite the fact that this mind wandering exemplifies a lack of vigilance.

Someone might object that this makes mind wandering indistinguishable from multi-tasking, as mind wandering and multitasking seem to be forms of plural pursuit. However, the difference is that when mind wandering, there is no commitment to maintain focus on any particular task or set of tasks, whereas there are such commitments when multi-tasking. In short, this means that the difference between multi-tasking and mind wandering comes down to the fact that multi-tasking can exemplify vigilance, whereas mind wandering does not.
One benefit of this account of mind wandering is that it is continuous with the self-control as vigilance account outlined in Chapter 3. Mind wandering, on this view, does not manifest a lack of control, but rather a redeployment of control that is distinct from focused, goal-directed cognition.

This view also accounts for instances of strategic mind wandering. For example, Seli et al. (2018-a) found that, in predictable experimental conditions, individuals concentrate their mind wandering in periods of low demand. Seli and colleagues take this to be evidence for intentional mind wandering, but the current framework provides an alternative interpretation. If mind wandering results from allocations of control, and allocations of control are sensitive to computations of opportunity cost, then we should expect to see mind wandering in conditions where plural pursuit has the lowest opportunity cost. Hence, the view outlined here explains how mind wandering is controlled without conceding that there is such a thing as intentional mind wandering.

Finally, this account solves the Puzzle of the Purposeful Wanderer using the conceptual tools of the TUT view. We can make sense of the fact that mind wandering is purposeless (and task-unrelated), while also explaining the numerous beneficial consequences of mind wandering. Additionally, we can see why mind wandering is useful for planning agents despite the fact that it manifests a lack of vigilance.

5.7 Conclusion

This chapter provides a sustained defense of the TUT view of mind wandering. Despite being the standard view of mind wandering, the theory has been challenged recently. I argued that these challenges exploit overly simplistic formulations of the TUT
view. When we draw out the connections between mind wandering and vigilance, we are left with a version of the TUT view that can respond to the various criticisms of it. I also outlined some difficulties of the alternatives to the TUT view. Finally, I outlined the way in which this account of mind wandering connects to the theory of self-control outlined in Chapter 3. This connection provides a solution to the Puzzle of the Purposeful Wanderer for the TUT view and helps to explain some recent findings that seem to support the existence of intentional mind wandering.
6.1 Introduction

This project investigates the nature and norms of vigilance. The primary aim is to understand how vigilance functions in human beings who act over time. Acting over time is a special ability that requires the use of planning and a special kind of self-control related to planning. This, I argue, is the background within which vigilance develops. In this concluding chapter, I summarize some of the main points of this project on vigilance. The bulk of this conclusion focuses on the novel framework for the empirical study of vigilance that emerges from the theory of vigilance outlined here. In particular, I discuss how to interpret vigilance tasks, major findings from the psychological study of vigilance, propose empirical predictions, and outline several areas of future research that could advance the study of vigilance.

6.2 Acting in Time

In Chapter 1, I introduced *The Problem of Scarce Information*. The problem arises from the way in which human beings pursue multiple goals over time. First, people aim to realize many goals. Each of these goals requires maintaining a certain level of
attentiveness to pursuing the goal. Human beings, however, have limited resources for attention, memory, control and other psychological processes related to goal-directed behavior. To compensate for these limited resources, human beings scatter their agency over time. Thus, people often have in mind less than the total amount of information needed to follow through on all of their goals at any given time. Scattering agency requires a mechanism for monitoring the environment to manage and regulate the flow of information in mind in order to efficiently pursue multiple goals over time. This mechanism, I claim, is vigilance.

We are now in a position to reformulate *The Problem of Scarce Information*. The reason for introducing the Problem is that we can characterize vigilance partly in terms of the design problem that it aims to solve. One issue with the original formulation of the Problem is that it relies on an information-processing model of cognition. While I endorse this model of cognition, there is a less theoretically loaded way to formulate the Problem that does not depend on information processing. I call this *The Problem of Acting in Time*. There are three components to this version of the Problem.

*Plurality*. Individuals typically pursue more than one goal at a time. This is in part due to the various pressures that shape early evolutionary history, where it is more efficient to build shelter, find food, and avoid predators all at once. However, social factors also influence plural goal pursuit. Meaningful projects often cannot be completed all at once. So, in order to pursue a variety of meaningful projects, one must balance pursuing a plurality of goals at the same time.

*Limitation*. There are limits to simultaneous goal pursuit. Some of these are purely anatomical. It is impossible to walk left and to walk right simultaneously in the same way that it is impossible to utter two distinct words simultaneously. However, some limits are psychological in kind. Attempts at multi-tasking often lead to performance decrements and mistakes. There are well-known disagreements over whether the sources of these psychological limits are metabolic, structural, or computational. But the simple fact remains that human beings are cognitively limited agents.
Scattered Agency. One way to compensate for Plurality under Limitation is to concatenate goals. This is not an efficient strategy. Many goals require coordinating with the world and with others in order to effectively pursue the goal. Thus, the appropriate time to pursue a goal is not fully dependent on one’s own decisions. One might start out doing something, only to realize that the opportunity for doing something else has arisen. A more efficient strategy is to scatter agency. One starts out doing something, switches to something else, and returns to the original endeavor at the right time. This allows for the possibility of engaging in a task even when one is not doing anything specifically task-related at present.

The problem of acting in time consists in coming up with an efficient design for scattered agency under conditions circumscribed by plurality and limitations. One way of putting the problem is this. Scattered agency is a strategy for plural goal pursuit in the light of limitations regarding simultaneous goal implementation. To do it successfully, it is not enough that one can hold in mind a goal-relevant task set, or that one can plan for the future. It requires, in addition, apportioning an action plan over time in a way that allows for satisfaction of a plurality of overarching goals. As such, it depends upon having capacities that go above and beyond those of agents who act at a time—regardless of whether the time for acting is now or later.

Vigilance, then, is a solution to The Problem of Acting in Time. This provides us with a way to initially characterize vigilance. Vigilance is a higher-order capacity that manages first-order psychological capacities to support scattered agency. This kind of capacity is constitutive of a distinctive form of self-control useful to agents that act over time and use planning to support temporally extended agency. In particular, it is a capacity that allows one to put on hold the pursuit of some goals, to focus on others, and to resume former pursuits at a later time. Vigilance, on this view, involves the management of attention, memory, etc. but is not reducible to any of these capacities or their conjunction.
6.3 Performance Breakdowns

We can also characterize vigilance in terms of the different performance breakdowns that appeals to vigilance can explain. Noting these performance breakdowns also provides an occasion to consider why vigilance cannot be reduced to the activities of first-order psychological capacities.

Consider again the planning errors mentioned in Chapter 1.

**Skipping.** Santiago is working in his office on campus. He needs to pick up a book from the library, which is on the way to his car. So he plans to pick up the book before heading home to do some grading in the evening. When walking to his car, Santiago ends up walking straight to his car, only remembering when he gets home that he wanted to pick up the book.

**Derailing.** Santiago wants to pick up a book at the library before heading home to do some grading. In the library, Santiago bumps into a colleague who makes several candid remarks about several proposed changes to the Faculty handbook that are up for review at the next Faculty senate meeting. The colleague’s remarks prompt Santiago to revisit some recent memos to brush up on the proposed changes. It is only next morning, before heading to work, that Santiago realizes he didn’t do any grading.

**Persevering.** Santiago wants to pick up a book at the library before heading home to do some grading. While in the library, Santiago starts perusing the volumes next to the book he needs, and ends up spending time skimming through several different books to see if they’re relevant to his research. After awhile, Santiago realizes that he’s spent too much time in the library and cannot do any grading this evening.

While these examples are simple, they all instantiate the same general pattern. There are a few things one wants to do, so the individual formulates a plan that would suffice to achieve the relevant goals. However, something goes wrong in the execution of the plan, which results in failing to do some of the things one had planned to do. The resulting activity is a mistake relative to one’s previously formed plan.

At first glance, explaining the occurrence of these performance breakdowns might seem easy. In **Skipping**, Santiago fails to remember his goal of getting the book from the library at the right time. Thus, his performance breakdown is due to a memory failure.
However, perhaps Santiago was deep in thought about how to present some ideas at a workshop he’s attending next week. In that case, Skipping looks like a failure of attention. In either case, it seems we can explain Skipping as a failure of memory or a failure of attention without appealing to anything like vigilance. The same goes for Derailing, and Persevering.

Vigilance, then, might not seem to add much to our cognitive economy. The function of vigilance seems reducible to attention, prospective memory, control, or some combination of these constructs. In Chapter 1, however, I argued that one cannot explain these mistakes as mere failures of attention, memory, control, or some combination thereof.

The anti-reductive argument relied on common patterns between each performance breakdown. In each case, some psychological capacity is engaged and functioning normally relative to the explicit task in which it is engaged. But the capacity is not deployed in the right way. This suggests that these failures do not arise from these capacities failing to perform their functions. Instead, these failures arise from the mismanagement or misallocation of these capacities. As vigilance manages lower-level capacities, the failures are really failures of vigilance.

The higher-order nature of vigilance generates the appearance of performance mistakes as failures of attention, memory, or some other first-order capacity. The activity of vigilance is to manage first-order capacities that mediate behavior. So, a failure of vigilance will also result in a corresponding failure of some first-order capacity. Hence, when vigilance fails there is a tendency to see only the attention or the memory failure.
However, it is important to keep in mind that these performance breakdowns are not mere attention or memory failures.

Importantly, however, this discussion shows that the activity of vigilance cannot be reduced to the functional profiles of attention, prospective memory, control, or some combination thereof. All of these capacities must be appropriately managed relative to task sets that enable plural goal pursuit. This management does not fall under the purview of these capacities or any set of them. A higher-order capacity is needed. There is a unique role for vigilance within the career of a scattered agent.

6.4 Planning and Vigilance

These dual characterizations of vigilance—from the Problem of Acting in Time and explanations of performance breakdowns—provide a conceptual framework for a theory of vigilance. This framework enables novel interpretations of previous empirical studies of vigilance and also makes several concrete empirical predictions about vigilance and the relationship between vigilance and other psychological capacities. Before getting to these, I want to first outline the main experimental paradigms used to study vigilance and consider some problems with current interpretations of results drawn from these paradigms.

The study of vigilance in psychology did not formally start until the early 1920s with Head’s (1923) research into the mechanisms for posture recognition and correction in the central nervous system. Head’s research did gain some theoretical traction (see Rosenfeld, 1924 and Blonski, 1928) and motivate some practical applications, particularly in reforms of driving tests in the UK and US. Miles and Vincent (1934) were
the first to propose that measures of vigilance (through iterated response time tests) be included in examinations for the driver’s license (Harry DeSilva was the driving force behind applied vigilance research in the US; see DeSilva, 1937, 1938, 1939; DeSilva, Claflin & Simons, 1938).

Some serious criticisms of Head’s framework, including his work on vigilance, surfaced in the late 1930s, culminating in Oldfield & Zangwill’s (1942) four-part critique of Head’s work and his major successors. In terms of vigilance research, the main criticism is that Head did not clearly identify the physiological basis of vigilance, and his predictions about the location of particular response mechanisms did not line up well with emerging neurobiological evidence (see Oldfield & Zangwill, 1942: 59-60).

Around this time, however, independent lines of research into vigilance were developing. During World War II, the British Royal Air Force began using radar technology to detect German vessels attacking Allied ships. Even though skilled operators used the radar equipment, people would consistently miss important signals. So, the British Royal Air Force wanted to understand why people were missing signals. Mackworth was charged with developing research to study the cause of these missed signals. To study the vigilance of radar operators, Mackworth developed the Clock Task. These studies defined the important questions that shaped the empirical study of vigilance through the late 20th century.

The radar operator’s job is quite difficult, as early radar equipment was hard to read. The signals are faint and not easily discriminable from false alarms. The Clock Task, on the other hand, is remarkably simple (when compared to its real-world counterpart). Participants sat at a desk facing a large clock whose hand typically moved
one second forward. Occasionally, the clock hand would move two positions (a *double jump*). The instructions were simple: when participants notice a double jump, they are to immediately report this (by pressing a button or making a mark on some paper). The double jumps would occur randomly over long intervals (~2-3 mins.) for long trials that lasted up to 2 hours (which matched the normal shift of a radar operator).

The major finding from Mackworth’s original round of research consisted in discovering a notable drop in performance shortly after starting the task (Mackworth, 1948). Reliably, after about 5 minutes, participants start missing double jumps. Then, they start missing more double jumps. Performance degrades consistently over the course of two hours. Figure 6.1 represents the average incidence of missed stimuli for all participants over the course of a two-hour Clock Task:

![Figure 6.1: Effects of a Two-Hour Watch in the Clock Task. The figure measures the average incidence of missed stimuli within half-hour blocks.](image)

182
Mackworth labeled this drop in performance the ‘vigilance decrement’, and subsequent research on vigilance has identified decrements in nearly every vigilance task (See, Howe, Warm, and Dember, 1995). After Mackworth’s research, researchers wanted to know what factors affected the onset and severity of the decrement (Davies & Parasuraman, 1982).

This work on vigilance was done in the heyday of behaviorism, and the theoretical assumptions of behaviorism cast a shadow over early interpretations of the decrement. Mackworth claimed that the decrement resulted from lowered arousal as a function of time on a monotonous task (Mackworth, 1956). Others thought that the decrement reflects low expected probability of stimulus occurrence (Baker, 1959). And others thought that people became habituated to the task, making task performance more automatic and correspondingly more susceptible to goal neglect and errors of omission (J.F. Mackworth, 1970). Various problems with these theories (cf. Broadbent, 1971) coupled with a shift away from behaviorism led people to adopt different explanations of the decrement.

Eventually, people considered the Clock Task a test of sustained attention and characterized vigilance as a capacity to sustain attention or as a kind of sustained attention (Mirsky et al., 1991; Oken, Salinsky, and Elsas, 2006; Posner & Petersen, 1990; Raz & Buhle, 2006; Robertson et al., 1996; Sturm & Willmes, 2001; Stuss et al., 1995; van Zomeren & Brouwer, 1994). These characterizations still represent the predominant views of vigilance (as noted in Chapter 2), and they inform two different hypotheses about the decrement. The Overload Hypothesis states that successfully performing a vigilance task requires sustained attention. The use of sustained attention consumes
limited information-processing resources. So, when people perform a vigilance task for a long period of time, these limited resources get depleted. When these resources are depleted, performance degrades (Caggiano & Parasuraman, 2004; Parasuraman, Warm, and Dember, 1987; Szalma et al., 1999; Warm, Dember, and Hancock, 1996, 1998). The Underload Hypothesis takes a different approach to explaining the decrement. On this view, vigilance tasks require sustained attention. But, participant interest and motivation dictates the direction or focus of attention. Over time, participants get bored with the vigilance task, so they start thinking about something else. When they stop thinking about the task, performance starts to degrade (Manly et al., 1999; Robertson et al., 1997; Robertson & O’Connell, 2010). While these two hypotheses are not directly incompatible (see Thomson, Besner, & Smilek, 2014), people tend to accept either one or the other explanation of the vigilance decrement.

There are some problems with each hypothesis, and these problems partly motivate the alternative interpretation of the vigilance decrement that I propose. First, let’s consider some findings that are curious if the Overload Hypothesis is true. One of the earliest findings on vigilance is that making double jumps occur at regular intervals (predictable to the participants) nearly eliminates the decrement (Deese, 1955; this effect has recently been replicated in Seli et al., 2018-a). Inserting warning cues prior to the occurrence of double jumps mitigates the decrement (Maclean et al., 2009). Increasing task engagement (or simply using ecologically valid stimuli) nearly eliminates the decrement (Barron et al., 2011). Likewise, including engaging, task-irrelevant stimuli exaggerates the decrement (Ossowski et al., 2011). These are curious because, in each case, participant interest seems to have a large impact on the severity of the decrement. If,
however, the Overload Hypothesis is true, then participant interest shouldn’t have this kind of impact on the decrement. In other words, if successfully performing a vigilance task consumes limited information-processing resources, then two participants that differ only in their level of interest in the task should not have dramatically different results (in the same way that two individuals of equal physical ability who differed only in their level of interest in a physical task should not have dramatically different results).

Relatedly, some easy tasks elicit vigilance decrements. So consumption of resources cannot explain performance deficits, because the onset of such deficits occurs well before one would expect if people were consuming a limited resource. Finally, proponents of the Overload Hypothesis have had difficulty identifying the limited resource consumed during vigilance tasks. There is no plausible biological candidate, and appeals to structural limitations or capacity limits seem either implausible or needlessly brute.

This might seem to be a straightforward argument for the Underload Hypothesis, especially given the difficulty that the Overload Hypothesis has in explaining the role that participant interest plays in task performance. While it is true that participant interest makes a difference for task performance, it would be a mistake to infer from this that task performance is solely a function of participant motivation and interest. One reason to resist this inference is that it fails to do justice to the real life cases mentioned earlier. Think back to *Skipping* or *Derailing*. Santiago doesn’t fail to get the book or fail to grade because he loses interest or motivation in obtaining the book or grading the exams. His frustration at failing in these endeavors provides some evidence for the continued interest. More vivid cases help to illustrate this point. Pilots sometimes forget to put down
landing gear before touching down. Parents sometimes forget their kids in the car. In neither case does the individual lose interest in the things they fail to do.

There are empirical reasons to resist this idea, as well. Mackworth (1948) found that inserting breaks intermittently during the task raises performance back to pre-decrement levels. An experimental group would, after one hour of performing the Clock Task, receive a phone call from a confederate in Mackworth’s lab. Participants reported a self-assessment of their performance and the confederate instructed the participants to do better in the next half-hour. Figure 6.2 depicts the result of this message:

Figure 6.2: Effect of the Telephone Message in the Clock Task. The dotted line reports measures from the control group. Note the significant improvement in detection in the experimental group from the second half-hour block to the third.
Participants in the experimental group performed nearly as well in the *third* half-hour as participants in the control group performed in the *first* half-hour (~85% average success rate). Proponents of the Overload hypothesis have pointed to the effectiveness of breaks as solid evidence for their view. Their line of thinking is relatively straightforward. Breaks allow participants to replenish limited information-processing resources. Proponents of the Underload hypothesis have difficulty explaining the effectiveness of breaks. Partly, this is due to the implausibility of positing that participants are discovering new motivation to perform the task.

The Underload hypothesis, then, does not provide any resources to explain the effectiveness of breaks merely by appealing to participant motivation. To furnish such an explanation, we need to provide a more nuanced view of the role that motivation plays in vigilance tasks (and in vigilance more generally).

Put yourself in the shoes of someone performing the Clock Task. The task is boring, perhaps one of the most boring used in cognitive psychology. The reason the task is boring is because there is nothing you need to do to perform the task successfully. You simply sit and watch for double jumps. Attention plays a role at the beginning of the task, when you’re unfamiliar with the particular clock, the rate of rotation, and the appearance of a double jump. Additionally, you’re likely performing the task in a psychology lab after having filled out some strange questionnaire and counting backwards from 400 in intervals of 7 while waving your arms. So you’re probably trying to figure out if there’s a catch or surprise or something else you’re supposed to look for. But no, it’s really all just about the double jumps. This all probably occurs in about 5 minutes. You’ve constructed a response policy, adjusted the policy as needed, and now there’s nothing left to do. That
is, when one is in the position of having a robust response policy, there is no further reason to continue actively attending to the task. This is where motivation and interest start to make a difference (in those cases where participants do not lose interest in the overarching task goals). Without any further reason to continue attending to the task, attention can be directed elsewhere, either to things that require more attention or just on things that are more pleasant to think about for the participant. Hence, participants do not always lose interest in the task. Sometimes, they are simply performing tasks where there is no reason for them to maintain attention on the task.

This way of thinking about the shift of attention during task performance suggests that people deploy attention strategically in performing these kinds of tasks. The intuitive phenomenology of task performance supports the claim that attention is allocated strategically. Additionally, the ‘economics’ of our psychological resources supports the idea that activating a particular allocation of attention is a matter of strategy. There are costs and benefits associated with shifting attention in one way or another. When one’s current task does not require attention (or, one’s allocation of attention to task performance fails to raise the probability of more efficient task performance), this raises the benefits of shifting attention elsewhere. Correspondingly, the costs of maintaining on-task focus increases.

One benefit of the idea that people deploy attention strategically is that this overlaps significantly with the underlying computational principles of other psychological capacities as well as recently elaborated general principles of cognitive regulation. The account, then, explains not only principles of strategic allocation of attention (itself an activity of vigilance), but also connects these principles to other
psychological research programs more generally. Over time, the hope is to reveal some
domain-general strategic allocation mechanism constitutive of temporally extended
agency.

Another psychological capacity that appears to be strategically deployed is drawn
from neurobiological studies of control allocation. Braver (2012) offers a general model
for thinking about the variable nature of control allocation drawn from two different
modes of acting on plans over time. Consider a situation where someone wants to get
milk at the store on the way home from work. One way to approach the situation is to
simply hold in mind the task representation of getting milk from the store to the exclusion
of most other things that occur during the day. This requires using working memory (to
hold the representation in mind), which has more strict capacity limits than other memory
systems (Anderson & Milson, 1989). Braver labels the process underlying this approach
proactive control. The benefits of using proactive control consist in high probability of
success. Working memory systems are closely aligned with sensorimotor preparation, so
holding a task representation in working memory leaves little space for interference or
goal neglect (Awh & Jonides, 2001; Magnussen, 2000; Magnussen & Greenlee, 1999;
Zaksas et al., 2001; Wilson, 2001). Conversely, holding multiple representations in
working memory generates severe interference effects (Fougnie & Marois, 2009;
Theeuwes et al., 2005). The cost of proactive control is the limited availability of
cognitive resources for other tasks. Hence, pervasive use of proactive control tends to be
highly inefficient (if not practically impossible) in dynamic, information-rich
environments. Another approach to planning to get milk is to store the task representation
in long-term memory. This representation is linked via associative mechanisms to some
cue (which could be either an exogenous event-based cue or an endogenous time-based cue) the detection of which triggers recall of the representation (Graf & Uttl, 2001). The benefit of this reactive form of control is that long-term memory is more capacious than working memory, so the storage of task representations in long-term memory frees up cognitive resources to pursue other tasks. The cost, however, amounts to a lower probability of task recall given the reliance on cue salience and attentional capture to trigger task switching. In a nutshell, then, proactive control relies on highly reliable but highly limited working memory, whereas reactive control relies on highly ‘cost-effective’ but more unreliable prospective memory.

The variable costs and benefits of different memory strategies informs sophisticated cost/benefit analyses about which kind of memory to use when performing a task over time. Some of these factors are well understood. For example, the expected task demands of future environments dictates whether people use more proactive or reactive forms of control (Lewis-Peacock, Cohen, & Norman, 2016). This makes intuitive sense. If you know that you will need to navigate a complicated environment, you don’t want to devote too many resources to holding in mind any particular task representation, especially if there’s some confidence that the right kinds of cues will be present to trigger memory recall. People also seem sensitive to expected benefits of planning when balanced against the level of effort needed to effectively plan (Kool, Gershman, & Cushman, 2018). Low benefits of planning on high-effort tasks correlated with more reliance on model-free reinforcement learning mechanisms, which corresponds to Braver’s model of reactive control. Finally, the predictability of stimuli relevant to alternative tasks plays a role in selecting a memory strategy (Seli et al., 2018-a). When
task environments are predictable, people can rely on more reactive forms of control to handle those tasks, freeing up more resources to use proactive control on the main task.

At a certain level of abstraction, these factors align with those that dictate deployments of attention. In both cases, there is a problem of balancing the costs and benefits of thinking hard about a particular task (Boureau, Sokol-Hessner, and Daw, 2016). Some tasks, valuable though they may be, can be performed reliably well without thinking very hard. When performing these tasks, people can shift resources toward other tasks. These tasks may be more pressing, or they may just be more pleasant (Inzlicht, Shenhav, and Olivola, 2018). Whatever the case may be, there appears to be significant overlap between the principles that govern strategy selection for control, attention, and memory. This is an important insight for understanding the organization of our psychology.

Thus, one way to interpret performance on the Clock Task is to see people shifting their resources away from task performance once they determine how to respond to the relevant stimuli and they realize the task can be performed well without much thought. In the language of other models, people shift from using proactive to reactive control. In light of this shift, there’s more emphasis on using prospective memory to perform the task well and a higher reliance on bottom-up attentional capture mechanisms. This actually generates a concrete empirical hypothesis for this interpretation of vigilance, which we’ll return to in the next section.

All of these considerations support an alternative theory of vigilance, which includes an alternative theory of the factors that influence the onset and severity of the decrement. The alternative theory is a Strategic Allocation theory of vigilance. On this
view, the activity of vigilance is to correctly deploy attention, memory, and control in a way that facilitates plural goal pursuit over time. In other words, vigilance is a higher-order capacity that manages lower-order capacities more directly implicated in goal-directed behavior.

This fits with how people perform on vigilance tasks and matches better with subjective reports of performance on vigilance tasks. Another aspect of Mackworth’s research shows that performance on a vigilance task does not degrade smoothly. Instead, performance degradation is dynamic. This reflects the strategic components to task engagement. Participants engage, disengage, and re-engage, as evidenced in the following results on performance indexed to the presentation of each double jump in Mackworth’s original experiment.
Figure 6.3: Effect of the Spacing of the Twelve Signals within Half-Hour Spells. This measures average incidence of missed stimuli for each distinct double jump. The markings at the top represent when the stimulus was presented, while each line plots performance over distinct half-hour blocks.

Participant performance peaks around the five-minute mark. This is consistent with claiming that participants quickly build and confirm a response policy, thereby sharply decreasing the value of continuing to focus on task. This implies a concrete hypothesis: after ~5 min., there should be a shift in neural activity from regions that subserve sustained attention and proactive control (such as tonic activity in the lateral prefrontal cortex) to regions that subserve reactive control and event-based prospective memory (e.g., transient activation of anterior cingulate cortex and lateral prefrontal cortex; see DePisapia & Braver, 2006).

Additionally, Mackworth’s results suggest that people re-engage with a task. No follow-up was conducted to determine the content of people’s thoughts prior or
subsequent to re-engagement (which appears to occur, on average, around 8 minutes and 15 minutes into the start of each half-hour trial). Additional studies could be conducted to evaluate the contents of probe-caught mind wandering around this time to see whether people are aware of sustained off-task thought that raises the value of returning focus to the clock task (however briefly). A recent study from Seli and colleagues (2018-b) indicates a strong correlation between the propensity to monitor mind wandering and the tendency to disengage from a focal task to endogenously monitor for a time-based PM cue. This suggests that there may be a computationally precise strategy that most people employ in performing a standard Clock Task.

Finally, the discussion of motivation suggests that strategic shifts in resource allocation are not simply the upshot of simple cost/benefit analyses of different rewards associated with tasks. People might shift resources away from performing a task that they value because there is no further reason to continue actively attending to the task. Hence, not only is the value of the target part of the cost/benefit analysis, but the expected value of allocating one’s resources to a task also plays a role.

In light of this, the full Strategic Allocation view claims that vigilance is a higher-order capacity that manages lower-order psychological capacities (memory, attention, control, etc.) through computing the values of various targets, the opportunity costs of pursuing those targets, the benefit of allocating resources toward pursuing each target, and the projected reward space that becomes available after pursuing these targets (see Chapter 3 for further discussion of these computational principles).
6.5 The Strategic Allocation Theory of Vigilance

This model of vigilance makes a number of empirical predictions, some of which were discussed in the previous section. In this section, I want to further highlight the concrete empirical predictions of the model and suggest different lines of research that might shed light on vigilance in planning agents. One of the theoretical benefits of the strategic allocation view is that it supports naturalistic interpretations of different phenomena associated with vigilance (the effectiveness of breaks, the feeling of effort, etc.) without resorting to the idea that the activity of vigilance consumes limited information-processing resources. Hence, a further goal of this section is to outline these theoretical benefits.

In the previous section, I discussed different principles that govern the strategic allocation of psychological resources in service of plural goal pursuit. This might make it seem like strategic allocation is the direct result of decision that people make, or that allocation is something that forms part of the phenomenology of acting over time. This is implausible, however. Most times, people probably strategically shift their psychological resources without making any specific decision about this allocation. Put yourself in the shoes of somebody performing the Clock Task. You probably start out, at the beginning, paying lots of attention to the clock for double jumps. After awhile, however, your eyes remain fixed on the clock, but your mind is elsewhere. You likely catch yourself after awhile and bring your mind back to the clock before drifting off again after a few minutes. These shifts represent strategic allocations, and yet there are no conscious decisions made. So strategic allocation need not be deliberate, conscious, or accessible at the personal-level.
Despite this, however, strategic allocation is not wholly sub-personal. To see this, we need to consider a story. In 1952, an ex-Naval officer by the name of Brian Shackel agreed to participate in one of Mackworth’s vigilance studies at Cambridge. Due to scheduling, however, Shackel had to participate in the study after a 48-hour sleep deprivation study (meaning, he walked directly from the sleep lab to Mackworth’s lab). Physically exhausted, a reasonable person might presume that Shackel failed miserably, missing more double jumps than any other participant. In fact, he performed the test perfectly, not missing a single double jump over the course of 2 hours. Here’s what happened: in an effort to stay awake, Shackel began pacing around the room while monitoring the clock. After a short amount of time, however, he noticed that the clock would double jump every time he reached 25 steps (part of the strategy for staying awake consisted in counting steps while pacing). A few minutes later, and Shackel realized that the clock was following a predictable pattern: 25 clicks followed by a double jump. So, Shackel shifted strategy. He began counting to 25 before switching back to the clock and noticing the double jump. Following through on this enabled him to perform the task perfectly (see Shackel, 1999 for Shackel’s own account of his experience with the Mackworth clock).

Shackel’s story raises many interesting questions, but the important part for the Strategic Allocation theory of vigilance is that Shackel deliberately reframed the Clock Task as a prospective memory task in order to compensate for his physical fatigue. This reflects the fact that vigilance tasks are highly similar to prospective memory tasks, often differing only in emphasis (prospective memory tasks emphasize performing the ongoing task, while vigilance tasks emphasize the prospective task). Emphasizing vigilance makes
participants better at detecting the target, but reaction times increase and accuracy for the ongoing task decreases. Emphasizing prospective memory reverses the pattern (see Brandimonte et al., 2001). Shackel’s performance is a function of deliberately construing the task differently to alter the demand effects. The lesson to draw from this is not that Shackel’s strategy is generally desirable or effective. The world had to cooperate quite a bit in order for Shackel to perform perfectly. But Shackel’s anecdote reveals that task construal can affect performance. This raises a question: how might task construal affect task performance more generally? Further, can people model plans in ways that differentially affect the cognitive load of executing the plan over time? While the Strategic Allocation theory does not make any concrete predictions about the cognitive load of mental models, Shackel’s story indicates that this is a worthwhile line of research to pursue.

The Strategic Allocation theory draws heavily on recent work studying the computational principles that underlie control allocation and memory strategies. According to these models, proactive control is closely aligned with sustained attention and working memory (Engle & Kane, 2004; Braver, Gray, & Burgess, 2007). As people shift resources, they rely on reactive control, associated with prospective memory, to perform the Clock Task. The effect of the factors that affect proactive/reactive trade-offs and working/prospective memory shifts are not well-understood in vigilance paradigms, and more work can be done trying to connect these different empirical programs. However, if the Strategic Allocation view is correct, then we would predict a shift shortly after beginning the Clock Task from proactive to reactive control. This should correlate with reliance on prospective rather than working memory as discussed in the previous
section. This also provides resources for interpreting why breaks are effective at returning performance on the Clock Task back to pre-decrement levels. If people engage the task reactively, then they rely on bottom-up stimulus capture to return attention to the task. This kind of stimulus capture relies on associative links between cues and the task representation that needs to be recalled from memory. The concrete hypothesis, then, is that as people switch to using more reactive control, they rely on prospective memory. In this case, the breaks serve to reinforce the link between cue and task, increasing the likelihood of successful stimulus capture subsequent to the break. In other words, the breaks don’t serve as a reminder about the task to be done; instead, they serve as a reminder *that there is a task to do*. Because the neural correlates of PM cue encoding, intention maintenance, cue detection, and recall are well understood, different neuroimaging techniques provide an objective way to measure whether this hypothesis is correct (Barban et al., 2019; Beck et al., 2014; Burgess et al., 2001; Kliegel et al., 2007; Okuda et al., 1998; Reynolds et al., 2009; Simons et al., 2006; West, 2011).

The Strategic Allocation view makes certain predictions about the parameters of task set maintenance. One feature of the Strategic Allocation view is that part of what determines the allocation of limited psychological resources to task performance is some computation about the expected value of allocating such limited resources as opposed to handling the task with more habit-based (“automatic”) information-processing channels. Sometimes, when an individual shifts tasks, this is because the previous task is such that the costs of staying on-task exceed the expected benefits. This implies novel explanations of why people allow themselves to be distracted in potentially volatile or dangerous environments. Consider, for example, texting and driving. People (let’s presume)
understand the risks of texting while driving. They value their lives and the cars they drive. The reason that people text while driving is not that they fail to appropriately value their lives or property; instead, people overestimate their ability to drive safely while not paying much attention (additionally, my guess is that people have a tendency to text and drive only in familiar environments, which familiarity adds to the overconfidence; that guess, however, is not a direct consequence of the Strategic Allocation view). One hypothesis, then, is that information campaigns directed at telling people about the dangers of distracted driving miss the point. People don’t fail to accurately assess the dangers; rather, they fail to appreciate how difficult driving is while distracted, even when driving a familiar route. Designing information campaigns designed to make the difficulty salient should (if the Strategic Allocation view is correct) reduce the rate of distracted driving.

One overarching issue related to task set maintenance is which factors affect goal maintenance of novel tasks within habitual routines. This is an important question for a theory of vigilance, especially given the kinds of performance breakdowns considered in Chapter 2. One pattern that emerged from different failures of vigilance is that failures of vigilance are more likely to occur when being vigilant entails successfully completing a novel task in the context of well-established routines. One example of this is the parent who normally goes to work, but due to unforeseen circumstances must take the kid to daycare rather than the parent that typically takes the kid. Or when someone has to bring something to work that they don’t normally bring. These situations leave one particularly susceptible to goal neglect, because of the strength of habitual information-processing pathways. Thus, an interesting line of research is to consider how alternative forms of
mental modeling might affect conditions of goals neglect. This is connected to the load of different mental models, but also depends on how complex plans conglomerate over time.

A promising experimental paradigm would test for interactions between plan factors associated with differential cognitive load (drawn from the experiments suggested earlier) and see how those different factors associate with other plans. For example, one experimental design could induce habitual routines for some task set over time before introducing a novel habit-incongruent task to be performed. Researchers could test how different task construals interact with different factors for habitual routines. Additionally, this design could incorporate certain value components to see how task performance interacts with motivation.

Finally, the Strategic Allocation theory of vigilance makes certain predictions about the relationship between vigilance and self-control (this relationship is also discussed in Chapter 3). The motivational aspect of the Strategic Allocation theory generates several concrete predictions about finding ego depletion effects in vigilance paradigms. The theory predicts that ego depletion results from allocating psychological resources to task performance when the costs of remaining on that task outweigh the benefits relative to the benefit/cost value of any other tasks the agent could focus on (this is meant to include a disengaged or plural pursuit stage where the agent does not allocate resources exclusively to any particular task). This suggests that it is possible to identify both ego depletion and reverse ego depletion effects. An ego depletion effect might emerge from a two-block experimental design. In Block 1, participants would perform an easy vigilance task (like an auditory vigilance task). In Block 2, participants would perform a standard self-control task (emotion regulation, working memory task, etc.).
Performance in Block 2 would be compared to performance in a control group. The hypothesis is that participants in the experimental condition would perform worse in Block 2 relative to controls despite performing an easy task in Block 1. The reverse ego depletion paradigm requires three blocks. In Block 1, participants would perform a standard vigilance or SART task for about 10 minutes (to induce a decrement). In Block 2, participants would play a game that requires complex spatial navigation or some other difficult but highly rewarding task. In Block 3, participants would again perform the vigilance or SART task. The hypothesis is that Block 2 would serve as a break despite being difficult and resource intensive. This implies that Block 3 performance would match initial Block 1 performance. This would support the alternative model of the interaction between vigilance and motivation suggested by the Strategic Allocation theory.

These experimental paradigms explore various dimensions of vigilance. Collectively, they investigate components related to monitoring, implementation, and maintenance. This aligns with the functional profile of vigilance outlined in Chapter 1. The aforementioned experimental proposals would test the various empirical predictions of the Strategic Allocation theory of vigilance outlined in this project. In particular, it targets various functional claims that distinguish the Strategic Allocation theory from various Underload and Overload theories of vigilance.

There is a robust empirical research program that falls out of the Strategic Allocation theory of vigilance. Future work could investigate these predictions to either confirm different aspects of the theory or identify areas where the theory could be refined. In the end, the overarching goal is to identify concrete strategies that people can
use to mitigate failures of vigilance. Developing such strategies, however, requires understanding different aspects of how vigilance functions in human agents situated in dynamic environments. One way to extend the experimental findings suggested above is to situate vigilance task performance in virtual three-dimensional environments. Because of the close connection between vigilance and prospective memory tasks, it would be possible to adapt spatial navigation tasks used in prospective memory paradigms to test predictions of the Strategic Allocation theory of vigilance.
BIBLIOGRAPHY


4. Ainsworth, S.E., Baumeister, R.G., and Boroshuk, J.E. Manuscript. “Glucose allocation during self-regulation is affected by cognitive assumptions and role motivations” (Florida State University).


264. Sinnott-Armstrong, W. 1985. “‘Ought to have’ and ‘could have’,” Analysis 45:1, 44-48.


224


301. Warm, J.S., Dember, W.N., and Hancock, P.A. 1996. “Vigilance and workload in automated systems,” In R. Parasuraman and M. Mouloua (eds.) Automation and


