Autonomy, Automaticity, and Attention: Why Empirical Research on Consciousness Matters to Autonomous Agency

By

Brandon D. C. Fenton

A Dissertation submitted to
the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

Graduate Program in Philosophy
York University
Toronto, Ontario

September 2014

Abstract

This dissertation addresses the question: what is personal autonomy? It begins by examining the main theoretical accounts of autonomous agency currently on offer. Although each of the available approaches faces significant criticism, I defend a revised internalist (and functionalist) account of autonomous agency which draws primarily upon the work of Frankfurt, Dworkin, and Bratman. Next, I show that recent work in scientific psychology (*viz.* research on automaticity) reveals new dangers for any account of autonomous agency (including my own newly revised internalist account). My response to the identified threat of automaticity draws upon research in the psychology of attention and, more extensively, on theorizing upon the unity of consciousness. I use a number of insights gleaned from these areas of research to then construct a more robust theoretical understanding of autonomous agency—one that addresses the worries generated by automaticity by proposing new and additional necessary and sufficient conditions for autonomy. What these new conditions entail is that individuals must possess a particular form of unified consciousness across time in order to have acted autonomously.

Dedication

For all those communities and individuals who have been unjustly stripped of their autonomy by the violent hands of external structures of power as well as those who continue to be unjustly deprived of their autonomy by such powers to this day.

Acknowledgements

First and foremost, I would like to express my greatest thanks and gratitude to my supervisor for this dissertation: Dr. Susan Dimock. Without her expertise, understanding, insight, and encouragement, this project would not have come together as it did, and the overall experience of writing it would not have been so enjoyable. I also want to thank the other members of my committee, Dr. Robert Myers & Dr. Kristin Andrews for their excellent suggestions and comments throughout the writing process. Next, I would like to thank all of the members of my oral examination committee starting with my external Dr. Michael Bratman whose work has been central to the development of this project, followed by the internal examiner Dr. Lawrence Harris and the Dean's representative, Dr. Verena Gottschling. Completing a doctoral degree is a long term project that requires other forms of social, emotional, and financial support as well. With respect to these forms of assistance, I would like to thank Nathan Monteith, Kirsten Fenton, and Amber Fenton for providing me with a place to stay for many years as I worked on completing this degree and for the inspiration and emotional support. I would also like to thank Greg Monteith and Shannon Hancocks for providing me with a place to stay in a time of need. Also, I would like to thank my father William Fenton, step-mother Kay Park, and mother Sabine Schenk for all of their support and encouragement both financial and otherwise. Finally, I would like to acknowledge the grant support that I received through the Ontario Graduate Scholarship program. The support provided by this award brought me peace of mind and provided me with the financial means to work almost exclusively on this dissertation.

Table of Contents

Introduction

From a pre-theoretical view-point, many of us may believe that we often act in the world in ways which are thoroughly intentional, willed, and autonomous. We may believe that we select for ourselves, for instance, things like which movies we want to go see, which meals we prefer to eat, which causes we would like to support, and what we would like to do with our lives. And how our behaviours unfold in the world may appear (for the most part) to be consistent with what we take to be our autonomous choices and actions. But against this foreground of our own pre-reflective understanding of ourselves (i.e. our unexamined self-conceptions) as relatively independent, autonomous agents, lies the backdrop of the world in which live and function—a world that is undeniably full of heteronomous influences upon our behaviour. This contrast between how free or unrestricted we might, in general, assume ourselves to be, and the concrete material constraints imposed by the world in which we live can sometimes come to collide in our experience, leaving us feeling powerless, helpless, and (self-)deceived. These kinds of disappointments can often spawn philosophical investigations into the sources of our troubles, confusions, and errors. In terms of the picture provided above, such a collision or confrontation of opposing factors may induce us to begin questioning our assumptions about just how free or independent we really are. That is to say, such disappointments may lead to our reconsidering the bounds of our own autonomy, or even just exactly what it means to be an autonomous agent.

The principal aim of this dissertation is to address the above question; namely, "What is autonomous agency?" It is a question that has generated a number of different theoretical responses in recent decades—the more prominent of which will be considered in the following chapter. However, each of the major theoretical models that have been advanced to make sense of personal autonomy has faced some significant critique and resistance. Moreover, much of the

philosophical theorizing upon autonomy would benefit from greater contact with certain important areas of more recent empirical discoveries about human psychological and behavioural functioning. This dissertation is thus, in part, an attempt to bridge the gap between our theoretical understanding of autonomy and certain aspects of our more recent scientific knowledge of human psychology. In particular, it is my contention that the recent findings in psychology on the phenomenon of automaticity reveal it to be a significant threat which radically challenges important aspects of our previous ways of thinking about autonomy.

The above points identify some of the central concerns that motivate this dissertation, and they also inform the general structural outline for how it is to proceed. The first thing to be considered will be the prominent models of autonomy currently on offer—these will be assessed in term of their independent plausibility, as well as in terms of their robustness in the face of automaticity. Second, a thorough elaboration of the ways in which automaticity has been studied as well as what those studies reveal about the phenomenon and the nature of the threat that it poses to autonomy will be considered. Finally, an examination of the research and resources available for responding to the problem of automaticity will then be examined in the service of finding some way to buffer an adequate theory of autonomy from the dangers presented by automaticity.

In Chapter 1, I begin by outlining the three dominant approaches to understanding autonomy, starting with the "responsiveness to reasons" view, followed by the "responsiveness to reasoning" view, before ending with the "coherence" view. I then provide some reasons in favour of adopting the coherence view of autonomy over the other two approaches since it appears to be most well suited to developing an understanding of self-governance that does justice to the idea that the power of self-governance is essentially a power *of the agent*. Later in

chapter one, I go on to cover three important models dedicated to establishing a coherentist view of autonomy. I begin by presenting Frankfurt's original proposal which is couched in terms of a hierarchical structure of desire and volition, followed by an important amendment to such an approach by Dworkin, before finally examining Bratman's more recent—though nevertheless deeply indebted to the earlier views of Frankfurt and Dworkin—temporally extended model. Following the examination of the three important coherentist proposals, I turn my attention to addressing the most significant objections to such an approach, and I argue that we need to reject the conceptualization of the sort of hierarchical structuring that was relied upon in each of the three coherence views covered. My proposal is that we can replace such a way of understanding the architecture of autonomy with one that instead focuses upon the fact that what underlies the hierarchical metaphor is simply a sort of coherence between desire and volition. In the final section of the chapter, I reveal that even with the most significant objections to the coherentist approach now addressed, there remains another significant threat to any model of autonomous agency that demands our attention. This is the threat that may be perceived by recognizing the extent to which our behaviours fall under the control of automatic processes. In the psychological literature, the sorts of behaviours that are automated in this way are known as instances of *automaticity*. The encounter with the notion of automaticity at the end of chapter 1 sets the stage for a deeper analysis of the research into the phenomenon in order to develop a more thorough understanding of the type of danger it presents to autonomous agency to begin in the following chapter.

In chapter 2, I begin by providing a basic outline of some of the research and central components of automaticity before distinguishing the notion of automaticity from the related legal notion of automatism (in order to avoid any confusion). After comparing and contrasting

these two related ideas, I then return to a more thorough and detailed analysis of automaticity, drawing upon a number of experimental studies in order to provide various concrete connections and access points from the empirical research to our understanding of the broader theory. The work done in these sections reveals a number of core features of automaticity uncovered both in the lab, and in real life studies established by self-reports (*viz.* in section 2.4 on actions not as planned). Later, because there is some significant behavioural and conceptual overlap between automaticity and habit, I examine the notion of habit and compare and contrast it with automaticity as well. I then go on to highlight the fact that a conscious dissociation from behaviour is at the very heart of automaticity before providing a number of examples designed to show precisely why automaticity is a serious threat to autonomy. With this greater understanding of automaticity finally in our possession, I suggest at the end of the chapter that we next begin to look into automaticity's experimental contrary—i.e. attention theory and research—in order to both sharpen our grasp of the contours of automaticity as well as determine whether or not we may therein find resources with which to buffer an adequate account of autonomy from the threat of automaticity.

Chapter 3 begins by noting that the field of attention research is not guided by a single accepted definition of what attention in fact is. This state of affairs has led theorists to develop a number of different and competing views of how best to both conceptualize as well as study attention. Most of the theoretical developments in the area have been shaped by the apparent shortcomings of earlier models. For this reason, I examine a number of the more prominent theoretical approaches to attention in the chronological order in which they appeared on the scene. Proceeding in this way allows us to both develop an appreciation for the obstacles faced by attention theory as it changed and grew over time as well as isolate questionable assumptions

made about attention in the early days of its study. The historical review of the theorizing and research on attention that is provided in this chapter begins with a preliminary on "divided attention" or "dual-task" research methods—since these types of studies are central to the majority of the work done in the field overall. This is followed by an examination of Broadbent's bottleneck theory of attention which was largely responsible for galvanizing the modern movement to study the phenomenon of attention. Next, I examine Kahneman's effort theory, followed by Wickens' multiple resource model, before closing out the review with the more recent "attention as selection-for-action" account. Following the review of the development of theorizing and research on attention, I then examine the relationship between attention and memory in order to point out the difficulties of disentangling these two related phenomena at the level of experimental design and interpretation of results. From there, I transition into an examination of a prominent multi-component model of memory advanced primarily by Baddeley that introduces the problematic notion of a "central-executive" that Baddeley initially characterizes as almost exclusively attentional in nature. One of the core problems with such a characterization being that it fails to connect with the most pertinent features of our ability to centrally control behaviour; namely, the volitional and intentional nature of personal agency. At the end of the chapter, I take stock of what the examined attention research and theorizing has to offer to our project of developing a model of autonomy that is resistant to the danger presented by automaticity. Finally, I conclude that although attention seems to play a part in mitigating a certain aspect of automaticity, it does not alone appear to provide us with the necessary tools to adequately confront the threat presented by automaticity to a theory of autonomy, and that we need to extend our search for an answer to that problem by considering the sorts of control that might accompany or underlie a particular form of unified consciousness.

Given that the research on attention did not provide us with sufficient resources to address the problem of automaticity, chapter 4 sets out to explore what further support for an adequate theory of autonomy can be gleaned from research on the unity of conscious experience. To begin this chapter, I provide a brief characterization of just what is meant by 'consciousness' and how it will be treated and understood throughout the chapter. From there, I examine the notion of the unity of consciousness, and explore some of the candidate types of relations (proposed within the literature on the topic) to potentially underlie such a form of unity. This provides us with a number of leads to consider as possible barriers to automaticity. With these details in hand, I return again to spell out in greater detail the precise ways in which instances of automaticity work to undermine autonomous agency. Following this more detailed analysis of the threat of automaticity, I then transition into my proposal for a solution to the problem of automaticity and the changes that this requires of our coherentist model of autonomy. Central to my revisionary proposal is the notion that autonomy requires a certain form of unified consciousness which I label "symmetrical unity." After the case for the modification to the internalist/coherentist approach to autonomy is made, I then consider and reply to several objections to the proposed additional requirements of autonomy. None of these objections, I argue, prove fatal to the suggested revisions to the adopted coherentist model of autonomy. Finally, to close out the chapter, I provide some concluding comments about what I take to be fruitful areas of continued future research.

Chapter 1

> There is a recurring theme that runs through most attempts to clarify the notion of autonomy. It is indicated by the etymology of the term: *autos* (self) and *nomos* (rule or law). The word was first applied to the Greek city-state. A city had *autonomia* when its citizens made their own laws, as opposed to being under the control of some conquering power.
>
> - Gerald Dworkin

1.0 Introduction

Although the genesis of the notion of autonomy may have occurred within the politico-judicial sphere as a label for city-states that sought to assert their sovereignty by legislatively conducting their own affairs independently from external rule, the term has since been applied to individual persons. As a personal attribute, the concept of autonomy[1] has been employed by philosophers and theoreticians in, as Gerald Dworkin says, "an exceedingly broad fashion" (1981, p. 54). It has been equated variously with notions of liberty, independence, dignity, self-governance, values, and freedom of the will (among other things).[2] It is thus appropriately taken to be what Ned Block (1995) has coined "a mongrel concept"—that is, a concept which lacks a singular common use and meaning, and yet, is treated as denoting something in particular.[3] More commonly, one might think of it as an *umbrella term*—that is, a term that covers a broad class of related concepts. Although we are right to treat the concept of autonomy in this way, there does appear to be a general consensus among theoreticians that a core understanding of the term involves the original idea of self-governance. The point of agreement seems to be that whatever

[1] From this point forward, unless otherwise stated, any mention of autonomy will refer to the personal aspect of the term.

[2] For some of the additional uses of autonomy see for example Nomy Arpaly's (2004) "Which Autonomy?"

[3] Similarly, Stefaan Cuypers (2000), in a nod to Wittgenstein, calls autonomy a "family resemblance" concept.

else autonomy might involve, it must in some way be concerned with how persons—like the city-states to which the term was first applied—individually govern themselves. In accordance with the general view, this dissertation will consider autonomy as a sort of self-governance.[4] About the only other point of widespread agreement between authors about autonomy is that it is generally taken to be something that it is good to have.

In addition to its being used to refer to a number of different things, autonomy is appealed to in many different intellectual and social contexts for a number of different purposes. For instance, it is frequently appealed to—and intimately connected to moral thinking—within the educational, medical, legal, and political spheres. This widespread application of the notion speaks to its importance to people generally as well as to social institutions and the interactions between individuals and these institutions. With respect to the educational domain, the notion of autonomy is involved in, for example, the adjudication of conflicts of interest and determining degrees of control over resources between educational institutions and governmental bodies.[5] In terms of its influence upon individuals within the educational system, it can be involved in things like curricular freedom and teacher leadership building, as well as, establishing policies which promote the development of independent learners capable of self-guided critical thought (as opposed to indoctrinated and dogmatic devotees).[6] In the medical profession, the notion of autonomy often plays an important role in things like the development of patient care practices,

[4] Admittedly, as Feinberg (1986) recognized, the notion of autonomy as 'self-governance' can be treated in various ways as well—for example, it can be thought of in terms of a *capacity* for self-control, or the *condition* of self-determination, or as a *sovereign right* to self-rule. But of these three views, what remains central to elucidating the concept of autonomy is an analysis of, as Christman suggests, "the actual condition of autonomy defined as a psychological ability to be self-governing" (1988, p. 110). It is this latter consideration articulated by Christman that will be the primary focus of this dissertation.

[5] This use is closer to its original politico-judicial sense.

[6] The philosophy of education and educational policy research has a rich history and vast body of literature dedicated to the topic of autonomy in education. To name a few notable contributions see: R. F. Dearden (1972) "Autonomy and education"; C. Winch (2006) "Education, autonomy and critical thinking"; J. White (1991) "Education and the good life: Autonomy, altruism, and the national curriculum"; and H. Siegel (1997) "Rationality redeemed?: Further dialogues on an educational ideal".

policy formation, and the understanding of informed consent.[7] In law, the concept of autonomy plays a fundamental role in establishing and protecting the liberty of both individuals and groups as well as establishing limitations upon the liberty of groups or persons when such liberty conflicts with the public interest.[8] It is also invoked in legal defenses where claims about a lack of personal autonomy (due to coercion or duress), may be deployed in an attempt to excuse defendants from criminal liability.[9] It is in part because a theory of autonomy has such far reaching interest and is often implicated in these and other important areas of human interaction and conflict resolution that advancing a more realistic[10] and refined account of personal autonomy is warranted. A better understanding of just what constitutes autonomous agency allows for more accurate and refined application of this concept to the various areas of interest mentioned above.

In addition to these social implications, a refined theory of autonomy is also beneficial for what it can tell us about ourselves as individual agents. The more clearly we can come to understand what is involved in acting autonomously, the more effectively we may be able develop the skills and habits that are supportive of greater degrees of self-control. Moreover, such clarity should also help us to be both better able to identify and more prepared to resist those influences that might undermine our autonomy. Therefore, it is not only in response to social concerns that developing a more refined and detailed account of autonomy is warranted,

[7] For a concise overview of the influence of autonomy in medical decision making see H. Brody (1985) "Autonomy revisited: progress in medical ethics: discussion paper". Also, for an excellent treatment of some of the multiple dimensions along which autonomy may be practically assessed in the medical setting see B. L. Miller (1981) "Autonomy & the refusal of lifesaving treatment". For more recent selections on considerations of autonomy in medical ethics see R. Kukla (2005) and K. Baeroe (2010) in references.

[8] See Sellars (2007).

[9] See Richards (1989).

[10] By my use of the term 'realistic' here, I mean an understanding of autonomy that captures a more detailed and nuanced rendering of real human psychological functioning than is currently available in the philosophical theorizing on the subject.

but also, because doing so has the potential to help us, as individual persons, to better navigate a world filled with heteronomous influences.

The next section provides a brief outline of some of the principal theoretical attempts to explain autonomy in addition to some reasons for thinking that only one of these views is on the right track and worth developing.

1.1 Three Approaches to Autonomy

Positive philosophical theorizing about autonomy has for the most part proceeded along three general lines of development. These general theoretical trajectories have been categorized by Sarah Buss (2008) as "responsiveness to reasons", "responsiveness to reasoning", and "coherentist" accounts respectively. Although any particular view that falls under the heading of one of these general accounts of autonomy may blend aspects of one general account with those from another, what sets an account apart from its competitors is its emphasis upon the centrality of certain features or conditions of autonomous agency. For "responsiveness to reasons" accounts[11] what is central is that an agent has the capacity to appreciate and in fact does consider the multitude of reasons there are—or at least a reasonably large set of these—for acting in any number of possible ways within a given situation. This type of account is considered 'externalist' since the reasons that an agent must be most responsive to have to do with the facts of the situation in which she finds herself.[12] The idea here is that if an agent is not sensitive to a fairly

[11] See Berofski (1995), Wolf (1990), and Fischer & Ravizza (1993; 1998) for insights into this view.

[12] Even though what an agent need be responsive to on this account can include facts about the agent's desires and interests, it concerns these primarily in terms of how they act as reasons among (or in relation to) the recognized

large number of reasons for and/or against various courses of action in a given situation, then she is not likely to do a good job of navigating that situation. And if she is unaware or unable to become aware of such reasons, there is a sense in which her power of agency is severely impoverished. In short, this account considers ignorance of or ineptitude with respect to identifying and working with the available reasons for action to be a barrier to effective and autonomous agency. One of the background worries that motivates this view is that a person who is insufficiently responsive to the reasons there are for action may be more likely to engage unthinkingly in activities that end up thwarting her own interests or aims. For example, an individual may wish to take the subway east in order to get home but fail to recognize that she is boarding the subway train from the westbound platform. She thus fails to take account of the existing *reason* (provided by the fact that she is boarding the train from the westbound platform) to correct her behaviour and bring it in-line with her goal of getting home. According to advocates of this view, it is these sorts of unwittingly self-interest undermining cases that make it difficult to treat such activity autonomous or self-governed.

Accounts that focus upon "responsiveness to reasoning", on the other hand, tend not to be as concerned with the reasons that are available to an agent in a given situation nor whether the beliefs that the agent has about these reasons be true or false. Instead, theorists who adopt this approach are concerned with an agent's ability to evaluate her motives in relation to her other beliefs and desires.[13] For them, what is most central to autonomous agency is this capacity for the calculative assessment of one's motives, which includes the ability to recognize the status of

---

facts about the external situation that she finds herself in—or more precisely, the reasons for action that those situational facts provide.

[13] This view is derived from the work of authors like Christman (1991; 1993) and Mele (1993). It is more commonly thought of as a historical account of autonomy in that the background beliefs, desires, and values against which new motives are weighed consist of preexisting or previously adopted positions. It is important to note, however, that for Christman at least, active endorsement of the formation of one's more stable or entrenched (i.e. historical) beliefs and values is not required. Rather, for him, it is enough that an agent did not resist or would not have resisted their development had she paid attention to them.

one's motives among the host of one's other beliefs and desires in light of this assessment. Furthermore, they contend that when this critical, evaluative process of practical reasoning is thwarted either by manipulation or indoctrination—i.e. where one's ability to effectively assess one's own motives in relation to one's other beliefs and desires is hindered—one cannot be considered to act autonomously. An important component of this view is that it allows for the agent to revise her previous beliefs and values in light of new information or experience (regardless of whether or not this sort of revision is uncommon or rarely actually occurs throughout the course of an individual's life). That is to say, the faculty of reasoning that theorists from this perspective regard as central to autonomous agency not only functions in terms of evaluating new motives in relation to views already held by the agent, it also works to grant the agent the ability to revise previously held beliefs and values. It thus permits that significant transformations may occur with respect to what an agent might take to be her core values. This view might appear to be more 'internalist' than the one previously mentioned, since it is not so much the considered reasons themselves that are most important, but rather, it is her ability to work with these reasons well. That is, it is her personal capacity for the evaluation and integration of reasons, and ability to grasp what follows from her calculative deliberations that matters most on this view. Nevertheless, the view can still be characterized as externalist since the type of reasoning at issue might itself be considered a sort of independent formal system (or culturally codified process) that an agent may dismiss if she so chooses.

Last on the list of the main positive theoretical accounts of autonomy is what Buss (2008) calls the 'coherentist' approach.[14] These types of accounts are more commonly treated as "higher-order" or "hierarchical" views but, for reasons that I will provide later (in section 1.3), I

[14] This approach to autonomous agency, advanced by philosophers like Harry Frankfurt, Gerald Dworkin, and Michael Bratman (see references), will be covered in greater detail later in this chapter.

prefer Buss' characterization of what is central to these views as a matter of coherence. On coherentist accounts of autonomous agency what is of primary importance is that an agent must approve of, or endorse, her motives and desires in order for the actions that result from them to be considered autonomous.[15] Coherentist accounts of autonomous agency are considered internalist because the kind of endorsement that is key on this view is a power of the agent herself. For them, what counts most is not so much the reasons that one may perceive to be pertinent, or the method by which one may come to various calculative and practical conclusions (these are externalist characteristics of the other two accounts previously outlined), but rather, it is the power of the agent to reflectively either endorse her desires and motives or reject them as unwelcome intrusions upon her mental life. Moreover, on this view, if an agent is powerless except to disown the desires and motives that lead her to perform some action, then there is a rather strong sense in which she is not autonomously engaged in that action. In other words, she is not capable of governing herself in such instances and her disavowal of the desires and motives that move her acts as a sign of protest and frustration with her impotence in these types of situations.

Although each of the above mentioned approaches has something to contribute to our knowledge of what it might mean to be autonomous, it seems that, insofar as we want to restrict our focus to the idea of 'self-governance', the coherentist proposal appears most well suited to attain the objective.[16] The case for the superiority of the coherentist proposal is made by the fact

[15] This kind of agential 'endorsement' is spoken of in various ways by different theorists. It may also be referred to as 'acceptance' of one's desires and motives, or 'putting one's weight behind' them, or 'taking a stand' with respect to them, or believing them to 'make sense' in light of the kind of person one is, or 'identifying' with them. In each case, what is important is an agent's attitude towards her own desires and motives. This view clearly involves a form of metacognition as would appear to also be the case for the responsiveness to reasoning view; however, the responsiveness to reasons view, at least *prima facie*, does not appear rely upon such a form of metacognition.

[16] That is not to say that the coherentist view doesn't face any substantial objections, it certainly does, but the most notable of these will be addressed later on in this chapter (in section 1.3).

that the 'reasons' available to an agent (even if these are exhaustive), and the process of 'reasoning' that she may make use of in a given instance are not essential to who she is as a deciding agent.[17] That is to say, they are not a part of her power to decide to act one way or another.[18] Rather, they merely inform her and help her to make good, or rational, or expedient, or prudential choices. A natural consequence of the externality of these elements is that she may reject them. And if an agent may reject the counsel of the reasons available or the process of reasoning she normally relies upon, and yet still act decisively and with an awareness that she is acting under her own power to act, then the process of reasoning and the reasons there are cannot be necessary for her to act autonomously. Perhaps most problematic about the two externalist proposals is that they tend to over-intellectualize autonomous agency—that is, they render it too rationalistic—by treating the reasons and reasoning processes as key factors responsible for doing the bulk of the work. In a related vein, philosopher J. David Velleman (1992) recognized, when analyzing the standard 'belief and desire' account of action that:

> In this [the standard] story, reasons [or, we might here add, 'a process of reasoning'] cause an intention, and an intention causes bodily movements, but nobody—that is, no person—*does* anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part. (p. 461)

My worry about the two externalist accounts of autonomy runs parallel to Velleman's concern with what he suggests is the standard account of action: the issue is essentially that the more that

[17] To borrow a line from one of Christman's footnotes: "…adding an 'external' rationality condition as a requirement of autonomy…effectively separates the property of autonomy from the actual decisions and judgments of real people" (1991, p. 9).

[18] Nagel (2003) makes a keen observation on this point. He claims, "When someone makes an autonomous choice such as whether to accept a job, and there are reasons on both sides of the issue, we are supposed to be able to explain what he did by pointing to his reasons for accepting it. But we could equally have explained his refusing the job…by referring to the reasons on the other side…[thus 'reasons' centered approaches] cannot explain why the person accepted the job for the reasons in favor instead of refusing it for the reasons against" (p. 234-235). It seems clear then that explaining the agent's decision will have to involve some reference to *his power* to decide.

is made about the importance of what are, for the most part, transient reasons and processes of reasoning that are dissociable from the agent's capacity to decide, the more the agent is sidelined from the picture of (in our case autonomous[19]) actions and agency. Nagel makes a related pronouncement: "The more completely the self is swallowed up in the circumstances of action, the less I have to act with" (2003, p.238). In our case, the "circumstances of action" amount to the formal features (e.g. the type and structure of the process of reasoning), and changing components (available reasons) of externalist views of autonomy. And the greater the emphasis upon these factors, the more the former of intentions and enactor of behaviours (i.e. the agent) is displaced or demoted.[20] It would appear then, because the objective is to get clear on what it means for an agent to be self-governing, that taking the focus off of agential power and placing it upon externalist features[21] is fundamentally wrongheaded.

The above is not to say that reasons at hand and reasoning processes will entirely fail to be of use to our understanding of autonomy—on the contrary, they will likely have a significant role to play[22]—rather, the suggestion is merely that we should be careful not lose sight of the

[19] I should here say something about the difference between basic agency or action *vs.* autonomous agency or autonomous action. The standard or basic view of action can be expressed along broadly Davidsonian lines, wherein one or more belief(s) and desire(s) provide the reasons which lead to the intention to perform some action. Even if we build a more robust conception of the agent into this view (as Velleman would argue, is needed), there are still ways in which an agent can perform actions without being entirely autonomous. For example, a mugger could point a gun at me and demand that I hand over my wallet—now, assuming that I desire to live through this experience and believe that not handing over my wallet could get me shot and possibly killed, this belief and desire may lead me to form the intention to hand over my wallet to the mugger, and to actually do so. In such a case, although I may act with intent, my actions are the result of coercion or duress rather than my own self-governance. That is to say, had the mugger not made such a demand while pointing a gun at me, I would not have handed him my wallet of my own volition. Put simply, actions (as standardly conceived) may be heteronomous, whereas autonomous actions are not.

[20] Nagel goes so far as to claim that "As the unchosen conditions of action are extended into the agent's makeup and psychological state by an expanded objectivity, they seem to engulf everything, and the area of freedom left to him shrinks to zero" (2003, p. 244). This seems to me to take things a little too far to the extreme. Nevertheless, I think that the worry that motivates such a view is a legitimate one.

[21] Or at least, in the case of forms of reasoning, features that may be externalized.

[22] This is due to the obvious fact that autonomous agents so often do engage in deliberative activities—weighing both the available reasons and the relationships between these and their beliefs and values—in the service of arriving at some acceptable and attainable end. In other words, although reasons and particular reasoning processes are inessential to acting autonomously, they are nevertheless typically what informs the agent's own power to decide.

fundamental importance of the power to decide that belongs to the agent alone independently of the situation in which she finds herself. And one way of failing at this task is by developing a highly formalized, predominantly externalist account of autonomy. In order to quell the worries of those who may think that shifting away from reasons and reasoning based approaches could leave us with an impoverished theoretical workspace, it bears mentioning that an additional benefit of the coherentist approach is that it can readily accommodate key features of "responsiveness to reasons" and "responsiveness to reasoning" accounts while still maintaining the centrality of the power of the agent. That is to say, there is nothing stopping an agent, on the coherentist account, from either considering a vast number of reasons for action within a given situation or, from evaluating her motives in terms of their fit with the other beliefs and values she holds. Instead, according to coherentists, one may engage in either one or both of these activities—the only difference being that, for them, autonomy is primarily a matter of the agent's providing assent to whatever reasons or motives are in the end acted upon. For the reasons mentioned above, this dissertation will focus primarily upon the coherentist view of autonomy.

However, before taking a look at some of the thoroughly developed coherentist proposals on offer, a caveat: although in this dissertation I will deal with notions of self-governance and aspects of unified consciousness that are endemic to personal identity, I will mostly steer clear of debates over the actual existence of a 'self' and what that might entail. One reason for avoiding the issue of personal identity (or the self) is that we may come to a functional understanding of autonomous agency that can accommodate various conceptions of the self and I take such a noncommittal stance on the subject to be a virtue of the approach here taken. Therefore, 'self-governance' should herein simply be read to mean (roughly) acknowledged personal authorship and control, by the agent, of the actions he or she commits. Put simply, the question that will be

considered is: what is autonomous agency? And not: what is the self? Also, the primary focus of this dissertation will center upon our understanding of autonomous agency rather than merely one-off autonomous actions[23]—the former being more global in scope and extended across time than the latter more local idea. This is not to say that the examples used will always refer to drawn-out or long-term courses of action. Instead, the provided examples will often deal with the short-term behaviours of an agent. However, these examples will be deployed in the service of understanding a more robust and general theoretical notion of autonomy than one that is concerned merely with singular and isolated actions.

One of the more recent and thoroughly developed coherentist accounts is Michael Bratman's planning theory of autonomous agency. The planning theory Bratman develops involves explicit reference to self-governing policies that have the authority to play the role of the agent in autonomous action. These policies are temporally extended and serve to support the motivational maintenance of certain behaviours. But before elaborating upon the planning theory in any great detail, it is important to note that the theory has been advanced by Bratman as an improvement to the hierarchical model[24] originally developed by Harry Frankfurt.[25] The hierarchical model proposed by Frankfurt situates the autonomous agent at a level above basic desires and drives—a placement that grants the agent the purview to either endorse or refrain

[23] There is a sense in which autonomous action appears impossible without a persisting autonomous agent, but I will not argue this point here.

[24] Until otherwise stated, I will preserve the original terminology and structuring of Frankfurt's model of autonomous agency, including talk of lower or 'first-order' desires and higher or 'second-order' desires and volitions, as well as the hierarchical picture that such a view paints. This will be done out of respect for the integrity of the view as it was first developed and introduced. Later, in section 1.3, I will argue that shifting our perspective of the view to one that treats it as, in a more basic way, simply encapsulating the notion of coherence will help to buffer the original account against its most common criticisms.

[25] According to Bratman, the planning theory he provides can avoid certain damaging criticisms of the sort leveled at Frankfurt's original model (e.g. the regress problem). See Bratman (2000, p. 34).

from endorsing a lower, first-order desire by way of what he calls a 'second-order volition'.[26] It is a person's ability to form second-order volitions that distinguishes them—with respect to the capacity for the kind of self-governance implied by autonomy—from other creatures according to Frankfurt.[27]

Because Bratman's project draws upon Frankfurt's earlier model, I will first spend some time clarifying central aspects of the original hierarchical model. This will be followed by a brief treatment of some potential amendments (provided by Dworkin) to such a view. I will then outline Bratman's planning theory with an emphasis upon that component of his views that he calls self-governing policies. Later, I will defend against some criticisms of the original Frankfurtian view as well as raise some concerns of my own with Bratman's attempted solution to these criticisms including a potentially serious problem for his particular account of autonomy.

1.2 Autonomy as Coherence

1.2.1 The Original Hierarchy

Frankfurt's seminal contribution to debates about autonomous agency was set forth in his paper "The Freedom of the Will and the Concept of a Person". There he distinguishes between the common understanding of personal freedom—namely, the ability to do what one wants—and

[26] It is important to note that in "The Freedom of the Will and the Concept of a Person" Frankfurt does not seem concerned about the possibility that one's autonomy may be undermined as a result of second-order volitions having been previously conditioned by, for example, various social influences throughout one's development. In response to these potential concerns, Dworkin has identified the need for what he calls 'procedural independence'; see his (1976), "Autonomy and Behaviour Control". More will be said about this potential amendment to an account of autonomy later in section 1.2.2.

[27] For example he claims, "It is my view that one essential difference between persons and other creatures is to be found in the structure of a person's will" (1971, p.12).

a higher-order reflective capacity to either adopt or reject one's initial or first-order desires and inclinations. It is this second-order reflective capacity that is central to Frankfurt's characterization of the person (or the agent as we are wont to say). Not only are people pushed and pulled by various first order desires and motives, but, says Frankfurt, "[We] are capable of wanting to be different, in [our] preferences and purposes, from what [we] are" (1971, p. 12). Thus, whereas first-order desires concern our wanting either to do or not to do some particular thing, our second-order reflective capacity allows us to form desires and attitudes about those very first-order desires; that is to say, we may have desires and attitudes that have other of our desires as their contents. For example, one may desire to yell at a telemarketer for having been called and woken-up in the early hours of the morning while also not wanting to have such a desire because it will result in a worsened attitude toward the rest of the day. Such ambivalence with respect to desires, it seems, is not an uncommon occurrence in the day-to-day experience of most people.

Another important notion in Frankfurt's account is that of the individual's will. For him, it is not just our ability to have various desires with other of our desires as their objects that alone defines our particular kind of agency, but also, that certain of our desires have the capacity to serve as what carry us "all the way to action" (1971, p. 14). He thus equates the will with "the notion of an *effective* desire" (1971, p. 14)[28]; that is, with a desire that is motivationally powerful enough to result in the attainment of it's object. But more important than this technical

[28] It should be noted, however, that equating the will with an effective desire, as Frankfurt does, seems to render the notion of *akrasia* or the "weakness of will" conceptually impossible. One way to keep open the conceptual space needed for such a notion would be to say instead that will is an effective desire that is the object of a (positive) second-order volition. This way, an effective desire that opposes one's second-order volition can be treated as weak willed. However, Frankfurt could simply maintain that weak-willed action is just action against one's better judgement. In that case, one's will would remain merely one's effective desire, and it would be weak when it is contrary to one's better judgement—however, relying upon the notion of one's 'better judgment' here may be too strongly dependent upon reasons to fit neatly within Frankfurt's internalist model. It is for this reason that I prefer the previously mentioned modification to the notion of the will as a way of making conceptual space for weakness of will.

specification of the will—and more centrally distinguishing of characteristically human agency—is what he calls, 'second-order volition'. To illustrate this latter notion (and the complex structure of human willing), Frankfurt first begins with the example of a physician who believes that he would be better suited to help his psychotherapy patients if he understood what their craving for a drug was like. We are invited to suppose that this belief leads the physician to form a second-order desire in favour of the first-order desire to take the drug. Now it may be true that the physician wants to have the desire to take the drug while altogether not wanting that desire to become effective. This leads to the somewhat awkward sounding statement by Frankfurt that, "…insofar as he now wants only to *want* to take it, and not to *take* it, there is nothing in what he now wants that would be satisfied by the drug itself" (1971, p. 15). In other words, that one has a certain second-order desire does not entail that one also has the relevant and compatible first-order desire. And when this is the case, the individual's second-order desiring does not amount to a second-order volition. However, in instances where the individual has both the first-order desire and the complimentary second-order desire that the first be effective in moving him to action, only then does the individual have a second-order volition according to Frankfurt. And, as mentioned, it is these second-order volitions that are the hallmark of the deliberate (and we might say autonomous) agency of persons for Frankfurt.

In contrast to the willful agent (i.e. one who exercises second-order volitions) lies what Frankfurt calls the 'wanton'. A wanton, he suggests, is an individual with no interest in his (or her) own will. For Frankfurt, the issue of wantonness is a matter of degree. That is to say, depending upon the frequency with which individuals fail to form a second-order volition, we may correspondingly attribute to them a greater or lesser amount of wantonness.[29] But this is not

[29] This point clearly establishes that Frankfurt's focus here is squarely set upon a protracted view of agency (i.e. the idea of the person over time) as opposed to merely local actions.

to say that the wanton is thoughtlessly engrossed with first-order desires, since some of these may conflict or be nullified even by deliberation. Indeed, Frankfurt suggests, "a wanton may possess and employ rational faculties of a high order. Nothing…implies that he cannot reason or that he cannot deliberate concerning how to do what he wants to do" (1971, p. 17).[30] What sets the wanton apart from other rational individuals is not that he lacks the capacity for reflection, but rather, that, in the end, he does not care by which first-order desire he is led to act: he is merely occupied with the strongest of his basic inclinations and desires.

To highlight the differences between the willful[31] and the wanton, Frankfurt provides the example of the unwilling and the wanton drug addicts. The unwilling addict experiences a deep tension between two competing first-order desires; namely, the desire for the drug and the opposing desire to resist taking it. But what defines the unwilling addict is that he forms the second-order volition to resist taking the drug. Regardless of whether he finally succumbs to the pull of the addiction, there is a sense in which, though the desires that move him belong to him alone, he may still regard the forces which result in his taking the drug as not his own. The wanton, on the other hand, although he may also contain the conflicting first-order desires to either take the drug or not to, is not concerned about the outcome of this conflict.[32] As we might expect, he is indifferent to such aspects of his mental life. But this indifference towards his involvement in the selection between opposing first-order desires means also that he will not

[30] Frankfurt would later reconsider the implications of this claim. See his "Identification and Wholeheartedness" (1987, p. 176).

[31] By use of this term I mean to express both the positive and negative aspects of second-order volitions (i.e. to either be willing or unwilling for one or another of one's first-order desires to be effective).

[32] It should be noted however, that forming a second-order volition does not require that one's first order desires be in a state of conflict. Indeed, on Frankfurt's view, one may also be a *willing* addict. That is to say, even though an agent may have an irresistible desire to take a drug at the level of the first-order, so long as the agent forms a second-order volition to take the drug—one might imagine, for example, that the agent enjoys the psychological effects that the drug produces and the social atmosphere of the drug user lifestyle—his taking it is done autonomously. Admittedly, some readers might find this facet of the view problematic. They may instead see the second-order backing of an *irresistible* first-order desire as something closer to acquiescence than autonomous agency but I will not defend Frankfurt's view from such worries here.

come to full satisfaction upon the overcoming of one of these desires in favour of the other since, he is not himself invested in the outcome of this conflict.

An important part of what it means to invest oneself in the outcome of lower order conflicts between desires is that one forms a second-order volition in favour of a particular one of them. Of this process, Frankfurt claims, an identification is made between the agent and one of his first-order desires. And this espousal of and identification with a certain desire means that there is a corresponding withdrawal from those desires that lie in opposition to it. These two notions of identification and withdrawal are paramount to a proper understanding of Frankfurt's characterization of the structure of human willing. Another noteworthy aspect of that structure is the sense in which an agent may be said to exercise 'freedom of the will'. In short, Frankfurt claims that, "It is in securing the conformity of his will to his second-order volitions, then, that a person [or agent] exercises freedom of the will" (1971, p. 20). Thus, it may be said of both the unwilling addict and the wanton addict that neither of them exercises freedom of the will with respect to their drug taking. However, though they are each not free in this regard, they are individually so for different reasons: the unwilling addict because he does not obtain the will that he wants, and the wanton addict because he has no volitions of the second-order concerning his conflicting first-order desires.

Perhaps the strongest and most persistent objection to the hierarchical model proposed by Frankfurt has to do with the lack of a firm ceiling to these higher-order reflections. In other words, there seems to be no limit to the potential conflicts between desires at ever higher levels. Moreover, there doesn't appear to be any non-arbitrary reason as to why one should halt one's ever higher reflective ascension along the graded ladder of desires.[33] One may thus reasonably

[33] This is the regress problem that Frankfurt's view is often charged with. It will be addressed later in section 1.3.

ask why it is the second level of this hierarchical structure that is to provide the privileged seat of agency. But this is not a problem of which Frankfurt was unaware, and he first sought to remedy it by appeal to the notion of identification. Of this process, he says, "When a person identifies himself *decisively* with one of his first-order desires, this commitment "resounds" throughout the potentially endless array of higher orders" (1971, p. 21). It matters not, he says, whether we here interpret him to mean that such an identification results in a series of ever higher confirming desires or if it simply means that the question of even higher orders of desires ceases to be of importance to the agent. Indeed, for all practical purposes, wherever the agent comes to an identification of this sort—one presumed to typically occur at the level of the second-order—is simply where the agent's power is alleged to be located. Nevertheless, many philosophers have found such a response to the problem to be unsatisfactory, and Frankfurt's later appeal to the notions of satisfaction[34] and necessary volitions[35] have not quelled the concerns with his account. That is not to say that the hierarchal story of human agency that he developed has not had a significant impact on the imaginations of philosophers and shaped much of the theorizing done in the field. Surely, it has been of no small import. Rather, it seems that whatever position the hierarchical structure of volition is to ultimately occupy in a thorough account of autonomous agency, unless an adequate solution to the above mentioned regress problem is provided, its role can only ever be a partial one.

[34] Frankfurt's use of the notion of 'satisfaction' refers to an overall state of tranquility with respect to one's mental economy. And this is a state of the agent that does not require any kind of acceptance at a higher order. See footnote 62 on page 38 of this chapter for greater detail. For the full account see Frankfurt's (1992) "The faintest passion".

[35] Necessary volitions, for Frankfurt, simultaneously constitute and constrain the character of an agent's willfulness. See his (1982) "The importance of what we care about" anthologized in his (1988) book by the same title and his (1993) "Autonomy, necessity, and love" anthologized in his (1999) book entitled "Necessity, volition, and love" for greater detail.

1.2.2 Dworkin's Amendment

The above mentioned regress problem is not the only concern that faces a hierarchical account of autonomous agency. Indeed, Gerald Dworkin, although also an advocate of a hierarchical approach to autonomy[36], has identified another concern for this type of account. According to Dworkin, there is a problem with simply taking a second-order endorsement of (or a positive attitude toward) a first-order desire or motive to amount to an agent's acting autonomously. The problem arises because second-order endorsements and attitudes can themselves be influenced in ways that we would normally take to undermine a person's self-governance. For instance, one may be the victim of deception or of coercive threats or incentives (or both), and these may have the effect of undermining or reversing the type of endorsements that an agent would have otherwise made had such compelling forces not been implicated in the process. In these cases, Dworkin suggests, "…a person will feel used, [and] will see herself as an instrument of another's will." (1988, p. 14). He also maintains that, "Her actions, although in one sense hers because she did them, are in another sense attributable to another" (1988, p. 14). They are attributable to another in these types of cases because the agent's second-order endorsements and performance of an action is here manipulated, or forcefully imposed upon the agent by another. Dworkin believes that these kinds of influences result in an agent producing involuntary behaviours and, when this kind of thing happens, it fractures the connections that regularly hold between an agent's actions and his or her character. The upshot of all this, according to Dworkin,

[36] Interestingly, both Dworkin and Frankfurt independently published hierarchical accounts of autonomy within a few months of each other, Dworkin publishing first the article "Acting Freely" in November of 1970, followed by Frankfurt's widely influential "Freedom of the Will and the Concept of a Person" published in January of 1971. According to Dworkin (personal communication, May 27, 2012), both he and Frankfurt developed the idea independently although they had each been in communication with Robert Nozick on the topic.

is that when an agent is the victim of deception or coercion, that agent's actions are not autonomous.[37]

Another more worrisome way of conceiving of the problem identified above by Dworkin, is to consider that not only might the kind of deception or coercion that undermines one's self-governance occur at a specific and isolated instance, but rather, it can occur over a period (or even over the course) of an agent's development, in a kind of insidious and manipulative way, to impart second-order tendencies to endorse certain motives. That is to say, such influences might taint the formation of an agent's preferences. This is the kind of concern that in part motivates Christman's historical account of autonomy previously mentioned (in footnote 13). To illustrate the kind of impediment such influences pose to autonomy, one might here imagine an individual who was raised in an isolated, strict, and deeply ascetic religious cult. Such an individual may have been conditioned over the years, by way of harsh punishments for disobedience, to forfeit various forms of pleasure seeking behaviours or self-fulfilling activities—such that, as an adult, this individual has, by external compulsion, come to unreflectively internalize a preference for the motive of personal restraint even in the absence of other cult members or any threat of punishment. Normally speaking, we would not consider preferences formed in this way to be expressive of the agent's autonomy since they more accurately reflect impositions or directives inculcated by the group (i.e. other cult leaders and members). In other words, preferences formed in this way reveal chiefly the control of the group upon the agent's desires and actions and not her own freely reflective control over herself.[38]

[37] The solution to this problem of second-order preference manipulation, according to Dworkin, is to propose a necessary condition for autonomy, one that he calls 'procedural independence'. This condition will be explained later in this section.

[38] Christman (2007, p. 21) proposes three primary conditions of autonomy—revised from his earlier 1991 draft—in order to block such examples from counting as instances of autonomous agency, but these will not be taken up here for three reasons. First, practically speaking, his conditions demand too much of the agent by way of accurate and comprehensive memory and ability to identify a number of often subtle preference inducing factors of

According to Dworkin, there are additional and related factors that subvert the free and unencumbered functioning of an agent's capacity to form second-order endorsements and preferences. Here Dworkin has in mind influences that "…keep the agent in ignorance of the true determinants of his behaviour…[and] which rely on causal influences of which the agent is not conscious…" (1976, p. 26). He provides the examples of subliminal motivation (if it were possible) and instances of induced cognitive dissonance[39] since, in such cases, agents are not cognizant of the real forces that control their actions.[40] These types of influences, according to Christman, can be made to "…force a person to (prefer to) do something" (1991, p. 3), and they also keep people unaware of the manipulation of their preferences. What is distinctive about these types of influences is that they serve to undermine the agent's ability to adequately reflect upon her first-order motives by keeping her unaware of the ways in which her reflections, preferences and endorsements are being controlled by external events. In such instances, an agent may believe herself to be acting autonomously[41]—since, her action, being as it is the result of a higher-order endorsement of a lower-order desire, is structurally identical to normal autonomous agency[42]—but she cannot be taken to be expressing genuine autonomy since her

---

one's potentially distant personal history. Also, it requires too much expertise with respect to an agent's ability to identify and adequately account for the equally obscure "reflection-distorting factors". Next, if an agent's second-order preferences were successfully (even though forcibly) ingrained by the group, then we have no reason to think that she would feel alienated from them, since, such preferences—regardless of having been implanted by way of external manipulation from the group—over time, infect by conditioning, or simply become part of her sense of her own character. This point is echoed by Thalberg who claims that, "…harshly conditioned adults and children…show none of the reluctance…found among "unwilling" addicts, alcoholics, and smokers" (1978, p. 222). And last, Dworkin's simpler requirement of procedural independence seems to be a sufficient means of resolving the identified problem.

[39] For an explanation of the theory of cognitive dissonance see Festinger (1957) or Festinger & Carlsmith (1959). For a modern treatment of the view see Harmon-Jones & Mills (1999).

[40] One might, for example, add things like being hypnotized or unknowingly having been made to ingest drugs to the list of these kinds of factors.

[41] At least, that is, until she discovers (if she ever does) the previously unknown sources that were responsible for shaping her second-order thoughts and subsequent behaviour.

[42] As Dimock notes, "…external influences can undermine a person's autonomy without thereby undermining the subjective conditions of autonomy (such as the ability to reflect upon her desires)" (1997, p. 84).

actions are being governed by forces of which she is entirely unaware and that are thus unaccounted for and unapproved of.

The previous three paragraphs brought to attention different but closely related factors which threaten to undermine an account of autonomy that relies primarily upon higher-order endorsements. To be sure, if the kind of higher-order attitudes and thinking involved—that is, the thought and attitudes occurring at the level presumed to be the locus of autonomous willing—can themselves be manipulated and controlled to such a degree that we are not prepared to grant that the actions that follow from them amount to autonomous ones, then mere higher-order endorsement alone cannot secure the autonomy of the agent. But we need not be overly alarmed by what these factors represent. Indeed, it would appear that these factors (e.g. coercion, indoctrination, induced cognitive dissonance, *et cetera*), remain threats to one's self-governance on any model of autonomous agency. Moreover, when we recognize certain influences to either undermine or to otherwise be impediments to genuine self-governance, we need only ensure that they be classified as such, and amend our model of autonomy to include a condition (or conditions) that guard against their intrusion upon our understanding of the concept[43]—and this is precisely what Dworkin does. The additional condition that is required to inoculate our hierarchical model of autonomy from these specific influences, according to Dworkin, is one that he labels "procedural independence".

According to an early formulation, Dworkin suggests that understanding procedural independence "…involves distinguishing those ways of influencing people's reflective and critical faculties which subvert them from those which promote and improve them" (1981, p. 61). Here we see that procedural independence is concerned with securing the freedom of an

[43] Indeed, as Christman suggests, "A full specification of what it means to be *self*-directed, in a manner that captures what it means to be autonomous, simply will include the sorts of factors (or the conditions for such factors) that must be *absent* for such self-direction to occur" (1988, p. 110).

agent's second-order functioning from various factors that impinge upon or disrupt it. Later, in a book length treatment of autonomy, Dworkin would add: "It involves distinguishing those influences such as hypnotic suggestion, manipulation, coercive persuasion, subliminal influence, and so forth, and doing so in a non ad hoc fashion" (1988, p. 18). The requirement of procedural independence, then, designates a class of influential factors that are generally known to undermine autonomous agency and to which the agent must not be subjected or else she will lack autonomy. Clearly, procedural independence picks out a negative state of affairs—that is to say, an agent can only exercise her autonomy *in absence of* such subversive influences. To state it differently, her ability to be self-governing depends upon a *freedom from* these types of encroachment upon her preferences and ability to reflect on first-order desires and motives.

Some might be concerned, however, that things like normal educational practices may in some cases fall into the above class of prohibitions[44] and that this would be a particularly unwelcome consequence of adopting the condition of procedural independence since one of the primary aims of regular education is to promote the development of autonomous citizens. One of the worries here is that even where the aim of educators is to develop a capacity for autonomy in young students, it may nevertheless be inevitable that they at some early point impart ideas, practices, and skills to these students in a way that mirrors almost identically what most would consider more insidious forms of indoctrination[45]—and this betrays a serious inconsistency between educational goals and educational practices.

[44] For a 'responsiveness to reasons' based disarming of this kind of worry see: Cuypers & Haji (2006).

[45] For example, educators might compel young learners to adopt certain beliefs about, say, the desirability of critical reflection upon reasons and motives before the students are capable of forming any well thought out or reflective assessment of such a belief for themselves. And such comportment—i.e. getting others to adopt beliefs before they can assess them for themselves—is in keeping with the behaviours of those who would seek to indoctrinate.

As far as one is concerned specifically with the idea of procedural independence, however, there appears to be a way of distinguishing the kind of pedagogy aimed at developing autonomous agents from that aimed at indoctrination, and of keeping the former style of pedagogy off the prohibited influences list while keeping the latter on it. The main point to recognize here is that what renders genuine indoctrination an item on the list of factors from which the autonomous agent must be free is that it undermines the agent's second-order functioning. It does this in part by instilling preferences in ways that are not obvious to the manipulated agent. Often, such preference inculcation is coupled with an admonition against being critical of these very preferences. It is because the people in these circumstances fail to recognize the ways in which they have had their preferences manipulated along with their (implanted) strong reluctance to question and reflect upon such preferences that they often remain helpless victims of these powerfully controlling factors.

With respect to the kinds of educational practices aimed at developing autonomous agents, on the other hand, while they may rely upon inculcating certain beliefs, this is done in the service of promoting the kind of reflectiveness and critical mindedness that will eventually enable the students to independently assess for themselves their own motives, reasons, and (most importantly) their second-order preferences. It provides to these students the tools that, once developed, will allow them to then review and evaluate the very means of their acquiring such abilities. Moreover, they are not in any way hindered from rejecting aspects—or even entire models—of the thoughts proposed to them before they were able to reflectively consider or evaluate them for themselves. And this is because, unlike with cases of indoctrination, educating for autonomy does not include any admonitions against self-critical reflection (it does quite the opposite actually). Instead, it provides students with the skill set and ability to re-consider any

portion or aspect of their lives and development. As such, educating for autonomy, it turns out, is thoroughly consistent with the goals of autonomous agency.

One of the things that becomes clear from the treatment of the worry handled above, is that Dworkin's characterization of the notion of procedural independence can help to classify influences that might at first glance appear to be either difficult to categorize or, that may seem to require potentially counter-intuitive classifications. The way to distinguish between those items that fall on the list—i.e. the list of influences from which the autonomous agent must be free—from those that might look like potential list members is straightforward: If the kind of influence in question is of the type that undermines, lessens, impedes, or subverts an agent's second-order functioning, it belongs on the list; whereas, if it supports, encourages, promotes, or otherwise empowers this kind of functioning, it does not belong on the list.

For the remainder of this dissertation, Dworkin's amendment to the hierarchical account of autonomy (i.e. the condition of procedural independence) will be treated as an additional and necessary condition of autonomous agency. Where it is not explicitly mentioned in the context of other components of autonomy, the reader is invited to treat it as a tacit rider that will ultimately be appended to the final account.

As previously mentioned, the hierarchical model has enjoyed a wide influence. One philosopher who has attempted to reconstruct an elaborate account of autonomous agency—one that preserves in large part the spirit of the hierarchical model advanced by both Dworkin and especially Frankfurt—is Michael Bratman. In his account, Bratman calls for the inclusion of an agent's temporally extended practices of planning and behavioural guidance by self-established policies. His model of autonomous agency is also one of the most inviting with respect to questions concerning the role of conscious unity in the guidance and motivation of agential

action over time, since his is the first of such models to devote significant resources to spelling-out just what all is involved in a temporally extended view of autonomy. The following section will examine Bratman's planning theory of intentional human action and what it has to say about autonomous agency.

1.2.3 Agency and Time: Bratman's Account of Autonomous Agency

According to Bratman (2000), there are three key elements that are central to the story of autonomous human agency: 1- that we are reflective; 2- that we make plans and use other related mental devices and strategies; and 3- that we conceive of our power of agency as taking place across extended periods of time. With respect to the first of these components (i.e. our reflective capacity), Bratman takes a broadly Frankfurtian approach. That is to say, he characterizes our reflective capacity in terms of the hierarchical structure of higher-order conative engagement with lower-order desires; thus, to this extent, his view mirrors the position described in section 1.2.1. However, he is keen to note that this account of reflectiveness is not without its problems[46]—for example, how the agent should come to 'identify' with or endorse a desire regardless of the particular level at which the desire is to be located. That is to say, as was pointed out in section 1.2.1, it remains unclear why we should accept that the definitive power of agency is to be taken as always located at this higher level.[47] Nevertheless, Bratman's only amendment to the view at this stage is designating weak and strong forms of reflectiveness—the

[46] One such problem, Bratman notes, is the difficulty inherent in reconciling agent causal explanations with our standard event causal understanding of the world. This problem will not be taken up here. For the purposes of this dissertation, however, it will be assumed that a compatibilist view of autonomous agency is a viable option.

[47] For some worries about under appreciating first-order desires and their behavioural cues see Friedman's (1986) "Autonomy and the Split-Level Self".

former denoting the capacity to have higher-order attitudes towards first-order desires, and the latter denoting the capacity of the agent to 'take a stand'[48] with respect to a particular first-order desire.

The second core aspect of autonomous agency, for Bratman, involves our use of both plans and policies to shape and guide our behaviour. As planning agents, Bratman suggests, our behaviour is more complex than simply acting from moment to moment in a way unconnected with our previously established goals and aims or their future fulfillment. Rather, he claims, we often act in response to various complex and hierarchically framed forward-looking plans.[49] Moreover, these plans are ostensibly responsible for integrating and ensuring the harmonious unfolding and functioning of our actions and activities over extended periods. That is not to say that one may not revise earlier plans in light of relevant new information that may come into one's possession—surely, this is an essential caveat of the view—but, says Bratman, "Prior plans have…a certain stability: there is, normally, a rational pressure not to reconsider and/or abandon a prior plan" (2000, p. 26).[50] That is to say, the instrumental reasoning standardly involved in the formation of plans serves not only to chart out the course of future behaviours but it also acts as a (defeasible) constraint upon reconsiderations of the initial objectives as formulated.

Whereas a plan typically involves a specific, if partial, program for guiding one's behaviour with respect to some end in view, policies, on the other hand, embody a more general type of commitment according to Bratman. For him, a policy concerns one's conformity to a set

[48] By his use of this phrase I presume Bratman is referring to the kind of agential identification characteristic of Frankfurtian second-order volitions.

[49] Many of these plans, according to Bratman, are only partial and need not be overarching 'life' plans. For example, they may involve a not entirely filled out plan to vacation in Cuba over the winter, rather than a dedicated life-shaping goal of becoming say, a politician (although, presumably, weightier life shaping plans are not to be excluded from this model).

[50] This pressure, according to Bratman, is due to the perceived need for means-end coherence incurred by the instrumental reason employed in the formation of the very plans in question. For greater detail, see his (1981), "Intention and Means-End Reasoning".

of behavioural guidelines about what to do given a certain situation that is likely (or has the potential) to be repeatedly encountered. For example, one may have a policy of always asking one's coworkers whether they had a pleasant weekend upon seeing them at work on Mondays, or of always checking the tire pressure before going on a lengthy road trip. When the role of policies and planning in our action is given due regard, Bratman believes, we extend our understanding of autonomous agency beyond a basic 'belief and desire' psychological rendering. Indeed, although these manners of structuring one's behaviour may be generally treated as kinds of pro attitudes, there remains a sense in which they are unlike regular desires. What distinguishes our intentional plans and policies from the ebb and flow of our everyday desires, according to Bratman, is that, beyond their motivational roles, "they are subject to distinctive rational norms of consistency, coherence, and stability" (2000, p. 27).[51] As mentioned, it is not that the reasons involved in establishing such norms cannot be challenged or defeated, but only that there is some resistance to such changes in light of the globally instrumental role played by these plans and policies.

It might appear that the two components of Bratman's view addressed so far are separate and unrelated aspects of how we express our autonomy; after all, one (our reflectiveness) picks out our ability to consider our immediate desires and whether or not we want any one of them to be effective, and the other (our planfulness) serves to orient our long-term behaviours. But one commonality between the two is that each element partakes of a hierarchical structuring[52]: the first, by way of a graded scale of desires; and the second, in terms of ends over the means that

[51] For a potential challenge to this distinction see Alfred Mele's discussion of occurrent and standing desires in his "Motivation and Agency" (2003, p. 30-33). There, standing desires appear to capture something of the stability that Bratman attributes solely to plans and policies. However, Mele does go on to claim that, "Having no explicit representational content, standing desires are not explicit attitudes. Rather, they are dispositions to have explicit attitudes of a certain kind" (p. 32)—and, if we consider standing desires to be mere dispositions rather than attitudes, it appears, Bratman's distinction may withstand the potential worry.

[52] That is, so long as one thinks of such a model in terms of the hierarchical imagery provided by the original account.

they instantiate. Nevertheless, despite this shared structural feature, it seems, there may be some people who do not take part in both kinds of mental activity. For example, it would appear that one could be a nonreflective individual and yet still carry out plans (this seems consistent with Frankfurt's construal of the wanton). However, for Bratman, the opposite does not hold—that is, one cannot be a 'strongly' reflective agent about a particular and immediate set of desires without at least viewing oneself as persisting to some degree into the future, and planning to remain on the side one has taken. This brings us to the question of our conception of our agentic power as something that is distributed across time.

The third component in Bratman's triad of core features of autonomous agency has to do with our understanding of ourselves as temporally persisting. When we engage in drawn out activities—e.g. plan a wedding—we conceive of ourselves as the same agent throughout the entire process.[53] That is to say, for each subtask that we may be engaged in performing, from setting appointments with various photographers, to selecting the floral arrangements or booking the honeymoon suite, we take ourselves to be one and the same agent all along—that is, as one who has embarked upon the project, has completed certain component objectives, and who expects to carry out several more before all is said and done. Of course, one may divvy up one's actions to correspond to the various subtasks as occurring at a particular time (Bratman calls this a 'time-slice' view of agency), but the majority of sane people do not normally think of themselves and their actions and involvement with the world in this way.[54] This is why emotions like pride, shame, and vengefulness (among others) make sense to us and form a coherent part of our understanding of the world and our place in it. We feel pride or shame for having

[53] Barring, of course, instances where those processes are interrupted by catastrophic life-changing or character shattering events (e.g. being forced by territorial wars to flee one's home and social milieu, or unexpectedly suffering a head trauma that leaves one noticeably brain damaged).

[54] Indeed, even our ability to appreciate music betrays a sense in which a fluid temporal persistence characterizes our experience of ourselves over time.

accomplished or failed to have accomplished some previously set task, project, or goal. And we feel vengeful because we view the agent (i.e. one's self) who was harmed in the past as identical with the one who presently feels pain.

When we come to appreciate the kind of coordinated temporal distribution of our thoughts, actions, and self-conceptions in this way, according to Bratman, we arrive at an important truth; namely, that our agency is, in a very full sense[55], a temporally extended phenomenon. To frame this notion of the temporal extension characteristic of our form of agency, Bratman draws upon a modern take on a broadly Lockean view. This Lockean view is grounded in the interconnected psychological ties between one's self and one's memories, one's future oriented intentions and their fulfillment, and the continuities between one's desires and their kin. With respect to the psychological ties formed by way of the agent's future oriented intentional activity, Bratman posits that it is a kind of active monitoring and regulation by the agent of her motivations that ensures that such ties exist. The danger with such a view is that it seems to push the agent back one step from the kinds of actions that are supposed to constitute her agency.[56] In order to address this concern, Bratman claims, "we will want to appeal to states and attitudes whose primary roles include the support of connections and continuities, which, on a broadly Lockean view, help constitute the identity of the agent over time" (2000, p. 31). This affords Bratman the license to accord agential authority to such attitudes, since, if we take these attitudes to support the functioning of relevant desires, we can plausibly claim that the agent endorses them. But this leaves us with the question of just what attitudes might play the kind of supporting role needed.

[55] By this I mean to convey a kind of deep persistence not captured by simple one-off actions.

[56] The worry here is that we are left with the idea of a little person inside the head watching and controlling from behind the scenes, separate from the cognitive processes taking place. This is problematic because Bratman's objective is to develop a naturalistic, and "nonhomuncular" view of autonomous agency.

To answer this question, Bratman suggests, we need to take another look at our planfulness and policies.

First, a brief digression: so far, it might appear that Bratman's account has done little to distinguish autonomous agency from a merely purposive theory of human action since there has been sparse mention of anything distinctly characteristic of autonomy (that is, beyond the higher order reflective capacity already provided by Frankfurt). But, contrary to this possible appearance, his view is very much concerned with a 'stronger' account of agency—one centered upon temporally extended self-governance—and we may therefore rightly claim that his is a view about autonomous agency.[57] It is, in particular, the notion of the self-governing policy that makes it clear that Bratman's view concerns autonomy.

Our planfulness and policies, as mentioned, are constitutive of the kinds of psychological interconnections and continuities that characterize our agency over time. And it is in part their very role to induce a certain stable integration and harmonious unfolding of relevant intentions and behaviours by way of the connections just noted. Moreover, if we combine this view of our planning and policies with the kind of weak reflection (i.e. higher-order desires and attitudes) mentioned earlier, we allow a new understanding to surface; namely, that some of our policies may in fact be higher-order policies. Indeed, we may reflectively come to form, for example, a policy to be more helpful to those who seem to be in need, or to try to be a more compassionate or conscientious person.[58] What distinguishes these types of higher-order policies from the ordinary policies one may have is that they entail the reflective desire that such policies be

[57] The core difference between the Frankfurtian and the Bratmanian approaches to the topic being that the former takes the seat of autonomous agency to be located at a higher order, while in the latter, the autonomous agent is taken to be grounded in Lockean psychological ties.

[58] Notice that these kinds of policies are more general in character than ordinary policies.

effective in governing one's actions.[59] Bratman labels these 'self-governing policies'. And it is this self-governance, in the form of self-established behaviour guiding and temporally extended policies that may provide an answer to the problem of agential endorsement (or 'strong' reflection) according to Bratman.[60] Indeed, he suggests, "the agent's reflective endorsement or rejection of a desire can be to a significant extent constituted by ways in which her self-governing policies are committed to treating that desire over time" (2000, p. 34). In other words, Bratman proposes that the agent be identified with her self-governing policies—that is, where these self-governing policies operate to support or impede the relevant functioning of a desire, we may likewise say that the agent herself has either endorsed or rejected the desire in question.[61]

Of course, the looming problem with respect to any strictly hierarchically structured rendering of autonomy is the possibility of psychological dissociation or estrangement from higher-order desires or, in this case, policies. This is rooted in the same regress problem that was mentioned at the end of section 1.2.1. Bratman's strategy here, for addressing the problem, draws

[59] One's ordinary policies may be shaped, for example, by following a command, or by conditioning or imitation (none of which necessarily implies reflective acceptance). However, for Bratman, in order for a policy to be considered "higher-order" it must be reflectively desired by the agent. On this picture, it is not that the policy itself has a hierarchical structure, but rather, that it is considered from the hierarchical perspective as the object of a reflective desire. Moreover, the examples of higher-order policies that Bratman provides appear to be less concerned with the role of recurrent situational triggers than regular policies.

[60] This is the point at which Bratman's view perhaps most significantly departs from the earlier Frankfurtian model. Although Bratman does this in order to avoid the regress problem, it seems to make things worse for his proposal. For one thing, like the externalist responsiveness to 'reasons' and 'reasoning' based approaches did before, this move seems to shift the focus away from the immediate power of the agent to a potentially distant instruction bearing and behaviour constraining cognitive element (i.e. the specialized type of policy he identifies). For another, as I will argue, it appears that the agent may dissociate herself from or disavow such a cognitive element similar to how she may ascend to a higher order of reflection on Frankfurt's original model. And if this is right, then the problem of agential endorsement still looms for Bratman. There will be more on this point in the pages to come.

[61] Bratman goes on to elaborate that one of the key ways in which self-governing policies operate to engage various desires is via governing the extent to which such desires are taken "as providing a justifying reason in motivationally efficacious practical reasoning" (2000, p. 39). For concerns about a circularity within this view, and his response, see his (2002) "Hierarchy, Circularity, and Double Reduction". This component of Bratman's view will not be given further attention here, in part because of its strongly rationalistic bent.

upon Frankfurt's notion of satisfaction.[62] However, satisfaction, he insists, is insufficient to handle the problem when framed solely in terms of hierarchical desires.[63] What is needed, he proposes, is a view of satisfaction that is connected in the appropriate ways to the temporally extended nature of our agency—that is, "to satisfaction with a self-governing *policy*" (2000, p. 34). He cautions, however, that such satisfaction must be understood to be flexible enough to allow some conflict or even violation of the policies in question, and yet firm enough to resist certain kinds of conflicts. What the proposed policy satisfaction ought to resist, for Bratman, is conflicts with other self-governing policies wherein one policy is challenged by another.[64] Nevertheless, it seems unlikely that this move will be able to block the worry raised above.[65]

The hierarchical theories first advance by Dworkin and Frankfurt, and more recently extended by Bratman, have received much critical attention. As noted, the most widespread objection to models of autonomy that rely upon hierarchically structured relationships between desires has to do with agential endorsement or identification. In short, the objection concerns the apparent open-endedness of such desire based hierarchies. That is to say, on these types of models, although an agent might appear to settle upon some second-order desire as 'the one'

[62] According to Frankfurt (1992), 'satisfaction' with one's higher-order desires is what blocks the infinite regress problem for his view. This satisfaction is characterized in part as an "…absence of restlessness or resistance" (1992, p. 12). He argues: "To be satisfied with something does not require that a person have any particular belief about it, nor any particular feeling or attitude or intention" (1992, p. 13). And this is important since, if satisfaction did require some kind of separate cognitive element, then the agent might become dissatisfied with that element as well, and dissatisfaction here would reintroduce the infinite regress. This is why Frankfurt claims that, "Satisfaction is a state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition" (1992, p. 13). Because satisfaction is given this negative characterization—that is, one of a lack of both resistance to or restlessness about one's second-order desires—questions about one's potential to be dissatisfied with one's complete state of being satisfied are seen not to make sense. And therefore, worries about an infinite regress are prevented from re-emerging.

[63] Bratman argues in his (1996) "Identification, Decision, and Treating as a Reason," that Frankfurt style 'satisfaction' with one's higher-order desires is indistinguishable from one's having yet to decide one way or another about whether to challenge such desires, and so it does not appear to be enough to anchor the agent's endorsement of or identification with them. According to Bratman, what is required, then, is something stronger than the negative construal of satisfaction as "the mere absence of motivation to "change things"" (1996, p. 7).

[64] There appears to be a parallel here with Susan Hurley's view of unified consciousness as something that cannot support mutually inconsistent contents. See her (1998), "Consciousness in Action".

[65] For more on this point, we will return to Bratman's view and a potentially serious problem for it (and for any account of autonomy) in section 1.4.

with which she identifies or endorses, there doesn't appear to be any principled reason why the agent shouldn't (or at least couldn't) continue to ascend to ever higher orders of reflective assessment. The threat incurred by such a potentially infinite regress of reflective desire evaluation is that it seems to show that there is no stable or consistent level at which the agent's endorsement is to be reached, and thus, the mere fact that an agent might act from a higher-order desire does not alone seem to be enough to ensure that the agent therefore acts autonomously.

In the next section, two responses to the regress problem will be provided. The first draws upon an aspect of both Frankfurt's and Dworkin's responses to the issue that directs attention to the practical constraints on typical instances of autonomous agency. The second reply involves reconsidering the conceptual framework by which we understand those models of autonomy that have been described in hierarchical terms.

1.3 Putting an End to the Regress Problem

As noted in the previous section, what is commonly known as "the regress problem" has proven to be a serious stumbling block for theorists committed to developing hierarchical accounts of autonomous agency. The following excerpt from Christman (1991) provides some perspective on the problem:

> Any account…that presupposes that the desires that move an agent are 'accepted' by her will invite an infinite regress of desires in the explanation of this acceptance. For either a desire descended to the agent without her awareness or approval…, or the agent was able to judge whether or not this desire was acceptable. If the latter is the case (as must be on hierarchical 'approval' models), then the judgment about the desire will have to be based on (other) desires of the agent. Then the question arises about these new desires and their being approved or not by the agent, from which flows the infinite regress of desires. (p. 8)

Now, to be fair to theorists who adopt a hierarchical approach to understanding autonomy, it is not obvious that the forming of an endorsement of a second-order desire always or necessarily entails reasoned "judgments" about such endorsements[66]; especially not judgments that need to be shaped by the agent's other desires.[67] It could simply be the case that these kinds of endorsements are secured as a result of the desires in question being congruent with, for example, the agent's sense of self—or at least with their not offending or disturbing one's self-conception.[68] It may therefore be advisable to abandon Christman's particular construal of the problem and instead focus upon a more widespread take on it.

The common understanding of the regress charge against hierarchical accounts of autonomy can be advanced rather straightforwardly, it seems, by posing a simple question; namely, as Thalberg puts it, "Why not go on to third-story or higher desires and volitions?" (1978, p. 219). Indeed, for many theorists, setting the seat of autonomous agency at the level of the second-order appears both arbitrary and unjustifiable. Skeptics of this approach to understanding autonomy, it seems, are not compelled to accept that there are any good or principled reasons for an agent to halt her continued ascent to ever higher levels of reflective desiring. And for them, the absence of such a compelling reason makes specifying the conditions

[66] Indeed, as Frankfurt suggests, "…the conformity of a person's will to his higher-order volitions may be far more thoughtless and spontaneous than this…[it may occur] without any explicit forethought and without any need for energetic self control" (1971, p. 22).

[67] Such a rendering of the problem might speak more to Christman's favoured personal history based solution to it than it does to the common take on it.

[68] Frankfurt's notion of satisfaction provides another possible alternative to the suggestion that there needs be an element of cognitive judgment directed at one's identifications or endorsements. This is not to say that the agent must be completely unaware of why she is comfortable endorsing a particular desire, but rather, it is only to say that deliberate acts of judgment about her endorsements are not necessarily required. Indeed, as Frankfurt claims, "…the essential non-occurrence [of satisfaction with one's endorsements] is neither deliberately contrived nor wantonly unselfconscious. It develops and prevails as an unmanaged consequence of the person's appreciation of his psychic condition" (1992, p. 13-14). Therefore, contrary to what Christman suggests, neither do one's second-order desires "descend" to one without awareness, nor does it require a cognitive act of judgment about one's endorsements of them on Frankfurt's account.

of autonomy, as primarily involving second-order desires and volitions, tantamount to trying to hit a moving target by always aiming in the exact same location. On some occasions, where an agent, upon reaching a second-order volition, in fact does abandon any further reflective consideration of her desiring and willing, the view may appear to have made a direct hit. That is, it might seem to explain all that there is to explain about the agent's autonomous actions. However, in other instances, wherein, for example, an agent's nagging doubts about her resolve to set out towards accomplishing some goal keep her second-guessing her endorsements and keep her climbing to ever higher-orders of reflection, there the weakness of the model is made apparent. This weakness is an inability to specify with any consistency both at what level the power of the agent is located and, why it should be located at any particular level.[69]

With respect to the first worry (i.e. the apparent inability of hierarchical accounts to specify a consistent locus of agential power), one response would be to suggest, along with Aristotle, that "we must not expect more precision than the subject matter admits" (*Nicomachean Ethics*, book 1, chap. 3). That is to say, that we should not expect that the power of autonomous agency always resides at a conceptually neat and tidy single level of reflection (*viz*. the second-order). Autonomous agents are complex and multifaceted creatures whose ability to reflectively consider their desires and volitions allows for considerable flexibility (i.e. indeterminateness) with respect to the level of reflection at which they may finally refrain from considering things any further. Therefore, in some instances, one's endorsements may very well be made at even higher-orders of reflective desiring than is the norm.[70] Nevertheless, and this is to address the second worry identified above, it may simply be the case that agents *typically* make the kinds of

[69] In other words, it is a question of what grants a given level of reflection its special status.
[70] Both Dworkin and Frankfurt accept this to be the case.

endorsements that are taken to be the hallmark of autonomous agency at the second-order of reflection. This take on the matter is consistent with the views of both Frankfurt and Dworkin.

Although there may not appear to be any established limit to one's ever higher level desiring on hierarchical models, for both Frankfurt and Dworkin, there does appear to be what might be called 'practical constraints' upon just how high an agent is typically inclined to reflectively ascend. Indeed, Frankfurt claims that, although "there is no theoretical limit to the length of the series of desires of higher and higher orders;…common sense and, perhaps, a saving fatigue prevents an individual from obsessively refusing to identify himself with any of his desires until he forms a desire of the next higher order" (1971, p. 21). Here, "common sense" and a "saving fatigue" may render apparent to an individual that continued ascension to ever higher orders of reflection upon her desires, at a certain point, loses its relevance. In other words, the continual second-guessing of one's endorsements or commitments eventually begins to look not so much like climbing to ever higher orders of engaged reflection, as it does merely vacillation with respect to the commitments one is considering making at lower-orders of reflection[71]; whereas, to form a genuine endorsement of a second-order (or higher) desire *just is* to put a stop to uncertainty and vacillation and stand by one's commitments.

Similarly to Frankfurt, Dworkin admits that an agent's reflections upon her desires may surpass those of the second-order, and he also believes that people are nevertheless inhibited from taking this ability to extremes. For instance, he claims, "it appears that for some agents, and some motivations, there is higher-order reflection [i.e. higher than the normal level]…[however,] as a matter of contingent fact human beings either do not, or perhaps cannot, carry on such iteration at great length" (1988, p. 19). And this is just to say that generally speaking (at least for

[71] That is to say that, at a certain point, an agent may come to recognize that continuing to a higher-order of reflection is likely only to take her one step further away from settling the matter at hand; a matter that begins with first-order desires and her relationship to them.

human beings according to Dworkin) people tend to be conservative with respect to the number of orders of reflection that they will ascend to in forming their endorsements. And this, it seems, should not come as a surprise since, practically speaking, there are limits to the benefits of continuing to engage in reflections of a higher-order.[72] Dworkin's suggestion that people may not be capable of continuing in this activity to any great length might indicate something like the inhibitory fatigue that Frankfurt mentions, or perhaps some other cognitive barrier to abstracting too far from their initial concerns for action.[73]

In addition to those personal reflection limiting factors mentioned above (i.e. fatigue, a diminishing sense of relevance, and disinterest), it would seem that, frequently, the demands of everyday life also institute practical constraints that inhibit excessive reflective tendencies. Often, in various social and other settings, a person's actions are subject to various constraints of expediency. If a friend asks one to go see a movie and one remains paralyzed by some obsessive compulsion to continue to reflect further at every point at which it seems a committed response is within reach (i.e. at every previous order of reflection), one will not likely have that friend for very long. Even admitting that one cannot make up one's mind is a better response to the social demands in this instance than the entirely socially awkward behaviour of remaining silently[74] locked in an endless chain of ever higher-order reflection upon the matter.

[72] That is to say that, at a certain point, perhaps around the fourth or fifth order of desire let us suppose, the value of the difference between the levels of desire begins to wane for the agent and it might not make any difference to her whether her final endorsement is settled at the level of the fourth or fifth order. Moreover, where this difference remains most pronounced, it seems, is between desires of the first and second order since, it is there that the difference in character of the desires is most clear: first-order desires concern one's potential actions in the world; whereas second-order desires concern only desires of the first order (i.e. desires of a different character than they are themselves). Beyond the first two levels of desires, however, every further level takes for its object a reflective desire of the same character.

[73] We might here imagine, for example, a general disinterest in continuing the activity eventually setting in—perhaps a disinterest brought about by the monotony of the cognitive exercise.

[74] Of course, one could vocalize each step in this process but that would hardly render the behaviour any less uncomfortable.

It seems then, that there is good reason to think that practical constraints, in the form of personal ability or interest and social demands for timeliness, set the tone for the degree to which an agent may typically engage in the kind of reflective activity that is characteristic of hierarchical models of autonomy. In any event, with respect to the standard degrees of reflective activity, empirical research can help to determine the cognitive norms of agents. And it seems reasonable to think that these norms, if empirically ascertained, would likely end up being on the conservative side as Frankfurt and Dworkin suppose.

Nevertheless, skeptics could maintain that the above talk of practical constraints notwithstanding, advocates of hierarchical views have failed to directly address either of the theoretical challenges that were advanced on their own terms. That is to say that, to point to what typically occurs with agents is not exactly to answer the questions about what stops their reflective ascension—or rather, why it should stop—and why the level at which they do stop is in some sense special.[75] Skeptics of such accounts might maintain that to answer these challenges head-on will require more than a mere listing of pragmatic constraints. And this may well be so, but, advocates of hierarchical accounts might be left with another option here—an option that could save them from having to address these challenges at all.

One way for advocates of a hierarchical approach to autonomy to dispense with the regress based challenges of the skeptic is to show that they were ill-conceived from the very beginning, and that when an adequate understanding of the proposed model is had, such challenges are seen to be misguided and to no longer apply. To demonstrate the error of these challenges will involve (as was mentioned at the end of section 1.2.3) taking another look at the conceptual framework used to explain hierarchical models of autonomy. If it can be shown that the regress based objections of skeptics derive from a faulty apprehension of the conceptual framework being used

[75] In other words, what renders such a level of endorsement the seat of the power of one's autonomy?

then it might also be shown that such challenges fail to amount to real problems for these types of models.

When undertaking to examine the models of autonomy that have been classically described in hierarchical terms, what ought to be noted from the very outset is that the construal of autonomous agency as relying upon a hierarchically composed desire structure is essentially a heuristic strategy.[76] It is to speak metaphorically about the mental lives of agents in a manner that is useful for keeping clear the various distinctions and conceptual issues that arise when considering, for example, the differences between the types of desires that one has and one's relation to those desires in the production of intentional action. Talk of various "orders" of reflective desiring can, in the end, be dispensed with at no apparent cost to the substantive claims made by the theory. Gary Watson, it seems, was onto something when he stated that "since second-order volitions are themselves simply desires, to add them to the context of conflict is just to increase the number of contenders; it is not to give a special place to any of those in contention" (1975, p. 119). And this is why (as mentioned in section 1.1), I contend, it is better to think of what have traditionally been thought of as hierarchical theories of autonomy not in terms of the structural layering of "orders" or "levels" of desires, but rather, in terms of coherence instead. But, before getting into this alternative conceptual schema, more needs to be said about the dangers of treating the original hierarchal perspective literally.

The ease with which the metaphors of "higher-orders" and "lower-orders" of desires fit into our conceptual schemas, it seems, has helped to render them rather inconspicuous, and has enabled them, for the most part, to evade a certain kind of critical recognition; namely, a recognition of the fact that they are nothing more than mere metaphors. And this is what appears to be lost on skeptics who make infinite regress based charges against hierarchical models of

[76] Credit for this insight is due to my supervisor for this dissertation, Dr. Susan Dimock.

autonomy. It seems that they have been seduced, perhaps by the facility with which such mental imagery fits into the overall rendering of agentic power on these models, to treat such hierarchical concepts in a literal manner, or to treat them as essential components of these views.[77] Moreover, by failing to treat such notions as what they in fact are (i.e. mere heuristic tools) they end up reifying the hierarchical structure of these types of accounts of autonomy. And once the hierarchical structuring itself is given this kind of (unmerited) footing, questions and problems which derive from that structure are allowed a way to take hold.[78] But, where the understanding is clear that the collection of hierarchical concepts deployed to describe autonomous agency is a mere aid to learning or way of facilitating our understanding of complex cases, such questions and problems should not arise. That is to say, there are no questions about, for instance, at what "level" or "order" of reflection that a principled reason to stop reflecting further should appear, nor is there any question about the "special significance" of any particular order of reflection once these "orders" are recognized to be nothing more than a manner of speaking about the relations between certain desires had by agents. And once we abandon this particular way of conceiving of the mechanics of autonomous agency, we can begin to see how a re-thinking of such a model along the lines of a coherence among certain of an individual's desires can provide us with a more apt rendering of what is actually taking place within the agent as well as provide us with an account that isn't vulnerable to regress based challenges.

[77] Granted, the terminology of hierarchy is ubiquitous within the literature and it is not commonly made explicit by theorists in this area that these concepts can be treated as purely metaphorical. Some advocates of hierarchical approaches to autonomy may even disagree with me on this point. Nevertheless, it seems clear to me that such a collection of concepts is inessential to the view, especially if one is concerned only with providing a functional account of autonomous agency (as is my goal here) and not one that is forced to make any metaphysical commitments.

[78] That includes problems about, for instance, on Friedman's account, the "autonomy-conferring status" (1986, p. 23) of reflective desiring, as well as the "ontological status" (p. 28-9) of the psychological process of identification. On a coherence view, as will become clear in the paragraphs to follow, there need be no special ontological status for any of the psychological processes taking place within the agent, nor does there need to be any special relationship of conferral between some part of the agent that is presumed to be already autonomous and some other part believed not to be.

To make the change from a model of autonomy that concentrates upon a hierarchical view of desires and volitions to one that centers upon the notion of coherence does not require any real restructuring of the model. All that is required is that the earlier descriptions and the hierarchy laden language that was made use of be replaced by language that emphasizes the coherence between one's different desires instead of the difference in ranking between them. An additional benefit of doing so is that it seems to provide a more accurate portrayal of autonomy since it is not the differences between psychological parts of the desiring agent that reveals her power of self-governance, but rather, it is that parts of an agent's mental life may *unite cohesively* to form a triumphant expression of her own will that reveals this power.

To begin to make this terminological shift, it seems natural to employ the term "raw desires" to stand in for what used to be called "first-order desires" or "lower-order desires" since the notion of a raw desire still captures something of the unrefined or immediate nature of the kinds of desires that one has about states of affairs or potential courses of action. Whereas, with respect to what were previously known as "second-order desires" or "higher-order desires," it would appear that we need not introduce any new terms since the notion of a "reflective desire" is already widely used in the literature to denote the kind of desires that are concerned with raw or other desires; therefore, in this case, we need simply abandon the talk of second- or higher-order desiring and instead continue to employ the notion of reflective desiring in their place. And the same goes for "second-order volitions," these may now simply be understood as "reflective volitions."[79]

[79] Alternatively, one may here be tempted to use Bratman's classification of weak-reflectiveness and strong-reflectiveness (the former to stand in for reflective desiring and the latter in place of reflective volitions); so long, that is, as one is careful not to reintroduce any of the hierarchical language used by Bratman in the characterization of these terms.

Importantly, on this new way of conceiving of things, these newly transposed terms do not signify any sort of implicit hierarchical ranking or ordering of desires; rather, they merely speak to the experiential character of the types of desires that they are.[80] Moreover, on a coherence view, nothing about the psychological processes taking place within, for example, the wanton or the willful agents (as mentioned in section 1.2.1) changes.[81] That is to say, just as before, the unwilling addict experiences a deep tension between two competing desires (i.e. the desire for the drug and the opposing desire to resist taking it) only this time, these desires are understood to be *raw desires* since they concern in a straightforward way the courses of action that are open to the agent. And what is now understood to define the unwilling addict is that he forms a *reflective desire* to resist taking the drug. Just as before, he may, in the end, succumb to the pull of the addiction. But there remains a sense in which, though the desires that move him are certainly his, he may still regard the forces which result in his taking the drug as not his own; that is, since they are not a function of the coherent and unified structure of his willfulness and reflective and raw desiring. Instead, what results in his succumbing to the drug represents an outsider to that coherent and crystallized psychological unity—it is a rogue impulse.[82] Moreover, the wanton, as before, will also contain conflicting raw desires to either take the drug or not to; yet, as before, he is not concerned about the outcome of this conflict. He remains indifferent to these aspects of his mental life. For him, there are no reflective desires or volitions, only the pull of raw desires

[80] The difference in character of the kinds of desires is a result of their different objects. Raw desires have as their objects actions or states of affairs that can be brought about through action, whereas reflective desires have as their objects either raw desires or other reflective desires.

[81] In fact, there are no psychological differences at all between the entire class of earlier hierarchical descriptions and the newly developed coherence view of those cases. All that changes, with respect to these examples, is our way of conceptualizing that mental activity and the emphasis that is given to structural coherence instead of hierarchical structure.

[82] One might alternatively say that such an agent is divided against himself, or that the lack of coherence between his actions on the one hand, and his reflective desires on the other, betray the fact that he is not operating in complete control of his own behaviours.

of various strengths. And his indifference towards which raw desires in the end move him to act means that he is simply not interested in his autonomy.

Recall that an important part of what it means to invest oneself in the outcome of conflicts between one's raw desires is that one forms a reflective volition in favour of a particular one of them. On the previous understanding of the model, Frankfurt suggested that what we are now calling reflective volition entailed an identification between the agent and what we now refer to as one of her raw desires. And that this espousal of and identification with a certain desire meant that there was a corresponding withdrawal from those desires that were in opposition to it. Indeed, these two notions of identification and withdrawal were seen to be paramount to a proper understanding of Frankfurt's characterization of the structure of human willing.

From the point of view of coherence, the notions of identification and withdrawal may continue to be a part of the story of autonomous agency but, they will need to be characterized in a different light. For instance, the notion of identification should no longer appear to be mysterious, as Thalberg (1978, p. 220) noted, nor should it refer to some additional psychological activity on the part of the agent, as Frankfurt (1992) warns, but rather, on the coherence view that I am suggesting, the identification of the agent with one of her raw desires is simply a property of the coherence between her reflective volition, and reflective and raw desires. For these psychological components to cohere *just is* for her to be identified with the one of her raw desires that is a part of that coherent unity as opposed to some other raw desire for which this coherence does not obtain. On this view, there is simply nothing more to consider when seeking to understand what it is that makes it the case that an agent feels or in fact is identified with one of her raw desires. It would make no sense to think, on the contrary, that she would instead identify with a stand-alone raw desire that lies in opposition to and is excluded

from her greater psychological unity since, to identify with a raw desire is simply to coherently reflectively desire and will it. And although this new way of thinking of the agent's identifying with some raw desire might make her withdrawal from the other conflicting raw desire seem to be more of a passive affair, the fact that a particular raw desire is left as a stand-alone desire is enough to show that the agent's focus is clearly elsewhere.

Now that the new terminological conventions have been outlined for what should, from this point onwards, be referred to as a coherence approach to autonomy, we can return to the hierarchy based objections to the previous way of describing things to see whether or not such objections continue to pose any problems on the revised approach. The first hierarchy based objection had to do with the apparent limitlessness with which an agent could continue to engage in reflective desiring of ever higher "orders" in the absence of (what seemed to be) any principled reason to stop. The second part of the objection, had to do with the special, essentially "already autonomous status" of the level of desiring at which the agent would reflectively climb no higher (regardless of what level that turned out to be).

To address these charges from the new coherence view of autonomous agency, we may start by pointing out that there is no reference to "orders" of reflective desiring on this way of seeing things; instead, there are merely reflective desires, and although these may be numerous (including the possibility of many of which that will have other reflective desires as their objects), there is no notion of rank-ordering between them, nor is there any special status accorded a reflective desire that is not itself reflectively considered (i.e. what would have been the "highest-order" desire, and the seat of one's autonomy on the former view). On this new coherence view, one's power to be autonomous does not rest with one's last, unconsidered reflective desire (nor the "order" or "level" at which that desire was formerly thought to be

situated); instead, the power of the autonomous agent is seen to be a function of the coherence between one's reflective volition and those reflective desires (regardless of how many of these there are) and some effective raw desire. That is to say that, the above stated coherence is what constitutes the agent's particular power to be self-governing. Once it is recognized that not only does the new understanding not admit of any rank-ordering between desires but that it also does not treat the final unconsidered reflective desire as the seat of the power of autonomy, it becomes clear that the old and persistent objections to what was originally thought of as a hierarchical approach are no longer troubling on the new view (i.e. since questions about reflective desire ascension and status are seen to no longer apply). Nevertheless, the benefit of considering the points pressed by these objections is that we have arrived at a revised account that appears to be a more adequate candidate view than its predecessor.

But even equipped with this new, more adequate understanding of autonomy, it seems we are still not entirely in the clear, for the kind of psychological dissociation that was implicit in the first aspect of the regress challenge—a dissociation suggested by the agent's ability to always withdraw and become one step removed from a former identification of hers—may continue, it would seem, to cause problems for even a coherence rendering of things since, such a potential for dissociation threatens to undermine the very stability of that coherence.

In the next section, we return to Bratman's temporally extended account of autonomous agency in order to reveal in just what way such a form of psychological dissociation may continue to be a problem even after we adopt the new coherence centered way of conceiving of the formerly hierarchical model.

1.4 Another Problem for Autonomous Agency

One of the positive contributions of Bratman's account to a theory of autonomy is that it draws our attention to the importance of the temporally extended nature of our autonomous agency. It does this primarily by way of explicating the character of the psychological connections that underlie those autonomous actions that take place across extended periods of time. Importantly, part of what defines the character of those temporally extended psychological connections is a form of coherence between various psychological components; namely, the agent's connection to her memories, her future oriented intentions and an awareness of their fulfillment, and the continuities between her plans, policies, and desires. The coherence between these elements can be seen to complement the other more temporally localized sort of psychological coherence mentioned in the previous section. Indeed, if we combined these two aspects of psychological coherence in the right way[83], the resulting temporally extended coherence view, it would appear, may provide us with the most robust and promising account of autonomous agency developed so far. Nevertheless, this kind of view may still face some problems.

One of the core concerns that I have with what is original in Bratman's account has to do with an element that it treats as a central component of autonomous agency; namely, the self-governing policy. My worry is inspired by some recent empirical work in psychology[84]—work summarized in an important paper by John Bargh & Tanya Chartrand entitled "The Unbearable

[83] That is, in a way that refrains from reintroducing hierarchical notions and that therefore, is not required to make accommodations to the challenges of infinite regress. In other words, what we would be considering here is not just the Bratmanian model of autonomy revamped with the new coherence terminology but, in addition, we would be abandoning the idea that satisfaction with a self-governing policy is required to block those regress charges.

[84] Insofar as Bratman wants to develop a naturalistic and 'nonhomuncular' view of autonomous agency, it would appear that his view needs to be responsive and answerable to the empirically derived worries I raise in this section.

Automaticity of Being". In the article, Bargh & Chartrand draw attention to the research on skill acquisition that is "focused on intentional, goal-directed processes that [become] more efficient over time and practice until they [can] operate without conscious guidance" (1999, p. 463). Drawing upon empirical evidence, Bargh & Chartrand construe the gradual automatization of such goal-directed processes as helpful in disburdening an individual's "limited conscious attentional capacity" (p. 464).[85] The development of the automatic functioning of such goal-directed processes, according to Bargh & Chartrand, is conditioned by the "frequent and consistent pairing of internal responses with external events" (p. 468). The story they provide is one in which the individual (or agent for our concerns) must first afford a significant amount of conscious attention in order to attain the behaviour relevant to the goal he or she has established; but as the behaviour is repeated the ties between it and the goal become reinforced in a way that, little by little, take less and less conscious effort, until eventually, the agent's conscious attention is no longer required and "drops out".[86]

The concern that such research elicits for the Bratmanian view, in part, has to do with the similarity of self-governing policies to the ultimately automated goal-directed processes empirically identified. Indeed, as one's self-governing policies become more entrenched, and more consistently result in the desired behaviours that they serve to implement, the more they too, it appears, may become unconscious and automatic processes.[87] In such instances, we may

[85] Bratman acknowledges this limited conscious capacity but does not identify the associated worry it involves for our self-governing polices—a worry I develop in the following paragraph. See his, "Intention and Personal Policies" (1989, p. 452).

[86] This phenomenon is often experienced by individuals when learning to play musical instruments. At first they must focus intently upon how they are manipulating their instrument in order to carry the tune, but over time, the music playing becomes automatic and the individual can play even while blindfolded or when attending to various other things like her audience or the sound of her own voice while singing along.

[87] It is important here to note that the 'dropping out' of conscious attention that concerns me is different from the issue of estrangement that Bratman addresses. The former involves an abandonment of conscious attention due to habitualization, while the latter is concerned with a withdrawal of agential endorsement. Thus, the agent may be

resist identifying the behaviours so induced as autonomous or as the result of autonomous agency, since the agent seems to no longer be actively involved in her own behaviours. For example, one may have the self-governing policy of always coming to the aid of others who appear to be in distress. But there is a danger in such a policy's becoming automatic. Let's say that one sees an individual being mugged and, due to this automated policy, yells at the mugger to leave his victim alone (let us here presume that such a policy became automated through having repeatedly defended ones classmates from the schoolyard verbal assaults of other children in childhood). The danger is not only that the mugger may now turn his aggression toward our helpful agent (a danger that may seriously threaten the agent's safety) but also that, due to the automatic issuing of this seemingly impulsive behaviour, our agent was not able to accord sufficient attention to the situation to revise his policy in favor of, for example, the more cautious and safer behaviour of calling the police.[88] And I don't think that the example just provided picks out an isolated occurrence. Rather, I agree with Bratman that we do often operate based on self-governing policies, but because many of these may come to be activated automatically—and because such an automated activation of behaviour seems capable of obtaining even in the absence of the kind of coherently unified psychological economy outlined in the latter part of section 1.3—I hesitate to call the behaviours they issue examples of autonomous agency. Of course, the policy may have been consciously and reflectively self-established, but insofar as the desires it facilitates result in behaviours that were not previously consciously attended to (i.e. in close temporal proximity to the issuing of the behaviours in question), I think it problematic to

---

unconscious of the policy that guides her behaviour and yet not estranged from it—the worry here is that, if the effective policy is unconscious, she may not endorse it either.

[88] Bratman, it appears, does have an answer to such a worry. For instance, he suggests, "in an emergency situation I do not reconsider or abandon my policy, I only block its application to the particular case" (1989, p. 456)—but this assumes that the effectiveness of the policy in question, is forever within the purview of the agent's reflective conscious attention. And not only is this assumption (in light of the empirical evidence) unwarranted, it also appears to threaten a vicious circularity (i.e. the agent's approval here resurfaces as something distinct from the self-governing policy that is supposed to be constitutive of the agent's endorsement of a desire).

attribute their origination to the agent's autonomous action rather than simply a non-autonomous behavioural script.

This brings us to the question of the nature of conscious involvement with a temporally extended view of autonomous agency. What seems clear is that it is not only the psychological ties that Bratman describes that are relevant to our understanding of autonomy. In addition, it is the subtle occurrent features of the psychological states involved in, and taking place across these connections that matter; and these warrant a more thorough treatment. One way of developing a more detailed understanding of what is required for a robust and adequate theory of personal autonomy—that is, in terms of an agent's conscious involvement in her ongoing autonomous actions—is to examine those instances wherein an agent's conscious involvement in her actions is either fragmented or altogether absent. Once we are clear on just what appears to be missing for one's agency to count as autonomous in such cases, we will then be better situated to account for what is positively needed to bolster our account of autonomy. In the following chapter, we will examine the empirical research on automaticity, along with several everyday examples of instances of apparently purposive behaviours that are nevertheless produced in this automatic fashion. After developing a more thorough understanding of just what automaticity is and why it is a problem for a temporally extended view of autonomous agency (or for any view autonomous agency for that matter), we will be better situated to recognize some of the positive requirements of an adequate theory of autonomy.

Chapter 2

2.0 Introduction

In the previous chapter, I outlined the three most common approaches to understanding personal autonomy. I then gave some reasons to think that what I (along with Buss) call a coherentist account is most well suited to developing an adequate view of autonomy. However, it was noticed that such accounts, having been traditionally framed in terms of hierarchical models of desiring and willing, were left open to certain regress based challenges that appeared to seriously undermine these sorts of approaches. After addressing the regress based objections to a coherentist model of autonomy—primarily by way of showing that the traditional hierarchical structuring of such approaches is inessential and heuristic—I drew attention to a different problem that would appear to be of concern to any theory of autonomy.[89] That problem was raised in the context of a challenge to Bratman's proposed temporally extended view of autonomous agency—a view that treats self-governing policies (specifically those with which one is satisfied) as grounding the agent's autonomous actions. In short, the worry was that instances of automaticity might reveal a further and under-recognized sense in which an agent can become alienated from her actions.

[89] It is important to note, however, that the problem that I will be developing in this chapter is of concern to the different theoretical approaches to autonomy for different reasons. For the 'responsiveness to reasons' and the 'responsiveness to reasoning' based approaches, it reveals that these theories lack the resources to distinguish instances of automaticity from normal instances of autonomous behaviour since there is nothing about the way that such sequences of behaviours are formed or come about that contrasts with what these views have to say about normal autonomous functioning. Moreover, neither of these views appears concerned with the agent's attention at the moment of action, so long as the stated action can in some way be traced back to the reasons available for it or the process of reasoning that underlies it, regardless of when such reasons or reasoning took place, even if those thought processes were temporally significantly removed from the issuance of the behaviours in question. With respect to coherentist approaches, it reveals a further sense in which an agent can become alienated from her own actions.

In this chapter, I will clarify the notion of automaticity, examine some relevant current empirical research on the topic, and develop several examples to highlight why these types of automatic behaviours amount to a problem for autonomy that deserves greater attention.

2.1 A First Pass at Automaticity

Earlier (in section 1.2.3), I drew upon Bargh and Chartrand's (1999) summary of the empirical research devoted to what is widely known in psychology as automaticity. In that section, I described automaticity as a sort of behavioural script that may become operative without the conscious awareness—or at least without the occurrent conscious intention—of the behaving agent. It was also suggested that such automatic forms of apparently purposive behaviours were often the result of the frequent and consistent coupling of an environmental or situational trigger of a certain character (i.e. an external event type)[90] along with a standardly employed intention and performance of a particular sort of action.[91] With respect to the automation of various complex and skillful actions, the typical process involved in achieving a degree of mastery was seen to begin with the agent devoting significant attentional resources to acquiring the said skill. Indeed, in order to master such skills, the agent would at first commonly

[90] Indeed, Bargh claims that, "All automaticity is conditional; it is dependent on the occurrence of some specific set of circumstances" (1989, p. 7).

[91] It is important to note that, from this point forward, I will not be treating the phrases 'sequences of behaviours' or 'series of behaviours' as equivalent to the notion of an 'action'. Actions are the kinds of things that are performed by agents on my view—they require the occurrent conscious intention of an agent in order to be what they are. Similarly, autonomous actions are the kinds of things that are performed by agents that are autonomous for at least a given period of time. A series or sequence of behaviours, on the other hand, can be performed by non-autonomous (in the personal sense) systems that are devoid of any subjective sort of intentionality. For instance, a robot puppy children's toy can walk, sit, bark, and wag its tail but it would be a stretch to treat this behaviour as the intentional output of an agent as normally conceived. The reason for this terminological distinction is to differentiate certain of the constituents of automaticity (these will be called series or sequences of behaviours) from those of presumably standard and autonomous forms of agency (these will be called actions or autonomous actions respectively).

be required to devote a significant amount of practice time to the skillful action to be learned[92]—and this would include intense cognitive focus and bodily effort in order to both recognize and correct performance errors as well as to simply entrain the appropriate and desired series of behavioural responses. However, as the agent's ability to execute the skillful activity gradually improved, it was recognized that the intense concentration and effort devoted to acquiring the new skill at the outset would also gradually become more relaxed. That is to say that, as an agent's ability to perform a sequence of complex or skillful behaviours improved over time, the degree of cognitive effort and attention to the details of the performance of the action in question would correspondingly decrease. Moreover, after having achieved a certain degree of mastery over the set of desired performance behaviours, a remarkable result of the repetition required throughout the entrainment period was seen to obtain. The result was that one may then display the ability to perform the said sequence of behaviours devoid of conscious oversight or initiation—that is to say, automatically.

Although rather remarkable, perhaps the above characterization of how we acquire the ability to perform various complex behaviours automatically isn't all that surprising. After all, most of us may recall the various stages of learning and skillfulness that we each had to pass through in order to become competent automobile drivers as well as the attentional freedom that accompanies the comfort and confidence accrued over several years of successful driving. But, the initial phases of learning how to drive a car can be highly demanding on one's attention to detail and they are often experienced as quite stressful by many. Indeed, beyond requiring a sort of hyper-vigilant awareness or attention to one's surroundings when in the early stages of being behind the wheel of an automobile, one is also likely to be rather nervous initially about the

[92] As Dijksterhuis, Chartrand, and Aarts claim, "…for goal implementation to become automatized one needs to practice the selection and execution of the means in the goal-relevant situation" (2007, p. 104).

potential consequences of making mistakes while driving; perhaps most worrisome is the thought that an error could lead to the serious injury (or death) of another driver and his or her passengers, or pedestrians, or one's self. In addition to the risk of injury, one may also fear the possibility of damaging the family (or one's own) car, or someone else's property and the personal liability that might result from causing such damage.

Following the initial practice stage, one might remember the intense nervousness and the pressure of expectation that one felt when taking the licensing road test. In my own case, I recall being very alert to my environment, to the details of what I was doing, and to my memory of the training that I underwent prior to taking the road test. I remained attentive to the instructions given to me, observant of the road signs and driving behaviour of the people sharing the road with me, and consciously vigilant with respect to keeping my focus directed at the present and relevant details of my surroundings (including my own actions). And, remembering my training, I was sure to make it obvious that I checked the mirrors as frequently as it was stated that I should in the pages of the driver's manual. What has been described in the above example corresponds to and captures something of the high demands on attention during the initial learning period of the acquisition of most any new skill.

Despite the initially high demands on attention, however, several months after being awarded my license to drive an automobile, I gradually became a more comfortable and confident driver, and I found that some of my attention was freed up for things like enjoying the music on the radio, or carrying on a conversation while driving (these were things that struck me as terribly distracting during the early learning period, and such distractions were likely to provoke certain procedural errors on my part). This gradual transition away from tenseness and hyper-alertness to a state of comfort and confidence in my own ability corresponds to the

relaxing of the attentional effort typical of the gradual automation of behaviour mentioned previously. However, what may be most worrisome for many drivers is when that state of conscious relaxation becomes so complete that it ends up leading to episodes of full-blown automaticity. Indeed, frightening as it can be for us to realize, most of us who have been driving for say, more than a year, can recall a time when we had reached a stop sign with no clear idea of how we navigated the previous several blocks of roadway because our conscious attention was directed elsewhere. What these latter types of experiences reveal to us is that we can carry out very complex and dynamic sequences of behaviours without consciously attending to what it is that we are in fact doing.[93] This view is supported by Bargh, who claims that: "Once activated,…automated skills can interact with the environment in a sophisticated way, taking in information relevant to the goal's purposes, and directing appropriate responses based on that information, without the need for conscious involvement in those responses" (1997, p. 29). When we become skilled enough, we can allow these sorts of deeply entrained but nevertheless subconscious behavioural scripts or guidelines to take over and free-up our ostensibly limited conscious attention span for other things. Though it may be rather sobering to find one's self 'coming to' (i.e. coming back to a sense of awareness of what one is presently doing) at a stop sign after having navigated the previous several blocks absent of conscious attention to one's

[93] And this includes behaviours that are mediated by the changes registered by subconscious perceptions within a changing environment (such as the changing landscape and signage that accompany driving from one point to another). Indeed, according to Glaser & Kihlstrom, "…the results of research provide clear evidence…of what might be called contrast effects under conditions where controlled [i.e. conscious] processing is precluded, thus suggesting an *automatic correction process* [italics added]" (2005, p. 176). The authors further suggest that "the evidence for automatic correction calls into question prevailing conceptions of unconscious processes as passive and reactive…" (2005, p. 176). Behaviours that issue automatically may not only turn out to be automatically moderated by subconscious perceptions of relevant environmental changes but by the goal that is the terminus of the automated sequence of behaviours as well. And it can be this goal or end which provides the purposive structure to the automated sequence of behaviours that is responsible for shaping the 'automatic correction process' cited above. This would mean that instances of automatic behaviours are not always defined in terms of rigidly proscribed scripts, but rather, that behavioural adjustments may sometimes also be made both automatically and on the fly.

behaviours, what such moments reveal to us is just how commonplace and uncontrolled instances of automaticity are.

As a first pass at providing an account of automaticity, then, we ought to take note of the following characteristics mentioned above: 1- Instances of automaticity are made up of sequences of behaviours that an individual performs without direct conscious oversight or initiation[94]; 2- Instances of automaticity may betray a purposive structuring but they are nevertheless typically unsupported by the occurrent or active attention of the agent[95]; 3- Complex forms of automatic behaviour typically require an entrainment period that includes a frequent and consistent pairing of an environmental or external event (or part thereof) with a particular internal response; 4- Given a sufficient entrainment period, one's behavioural responses to familiar situations or certain sorts of environmental triggers can issue independent of conscious intention[96]; 5- Instances of automaticity often have the effect of freeing up one's active conscious attentional capacity for other things. As we progress through the relevant aspects of the research focused on automaticity, as well as several examples of automatic behaviours issuing in various everyday human activities in the remainder of this chapter, we may need to refine our understanding of these component characteristics somewhat. However, for the time being, we may take the above mentioned characteristics to be central to those aspects of automaticity that are of concern to an adequate theory of autonomy.

[94] Strictly speaking, one may consciously initiate a sequence of behaviours that, after having been started, one's conscious attention then recedes from (see: Bargh, 1997). However, many of our automatic behaviours are also initiated unconsciously. Indeed, Bargh suggests that research leads us to conclude that "...behavioral and cognitive goals can be directly activated by the environment without conscious choice or awareness of the activation" (1997, p. 47).

[95] Instances of automaticity often issue without the support of the agent's occurrent intention as well.

[96] As Bargh (2005) notes, "...evidence demonstrate(s) that action tendencies can be activated and triggered independently and in the absence of the individual's conscious choice or awareness of those causal triggers" (p. 38).

Before exploring the notion of automaticity any further, however, I will first take a moment to distinguish it from the related concept of automatism—a term of art that is standardly deployed in the field of law.

2.2 Automaticity and Automatism

Now that we have an appreciation of some of the central aspects of the sorts of behaviours typically classified by psychologists as instances of automaticity, we would do well to get clear on how these differ from, as well as potentially overlap with, those sorts of behaviours that fall under the label of 'automatism' within legal contexts.

As noted, the word 'automatism' is a legal term of art that has its application within the Canadian criminal justice system. In strict legal terms, automatism may be either *insane* or *non-insane*, the former denoting a mental disorder, while the latter simply confirms the absence of a conscious, voluntary act. The effect of a successful defence of automatism is the absolute exculpation, or acquittal, of a criminal defendant from legal guilt. The portion of Canada's Criminal Code that is of relevance here is Section 16 which deals with the defense of mental disorder.[97] In subsection (1) of section 16, it is stated that: "No person is criminally responsible for an act committed or an omission made while suffering from a mental disorder that rendered the person incapable of appreciating the nature and quality of the act or omission or of knowing that it was wrong." Although the notion of automatism is not explicitly codified within the above statement, it has made its way into criminal jurisprudence (or case law) as a consideration

[97] Interestingly, as Kalant (1996) notes: "The term automatism does not appear in either the ICD-10 (International Classification of Diseases, 10th edition, World Health Organization) or the DSM-IV (Diagnostic and Statistical Manual, 4th edition, American Psychiatric Association)" (p. 634), although it does appear in medical journals and text books. Nevertheless, Healy (2000) claims that, "The initial presumption that automatism results from mental disorder is far removed from medical understanding of the subject" (p. 87).

relevant to section 16.[98] Because of the emergence and importance of the notion of automatism as a consideration relevant to section 16 in the case law, in 1993 a review of section 16 was proposed by way of a White Paper which aimed to establish the verdict of not criminally responsible due to automatism. In this proposal, automatism was defined as "a state of unconsciousness that renders a person incapable of consciously controlling their behaviour while in that state."[99] Nevertheless, the review committee determined (with governmental endorsement) that both the definition of automatism and its legal application be left to the courts.[100]

Clearly, the above definition of automatism bears some resemblance to a certain characteristic of automaticity that was identified in the previous section; namely, in terms of the lack of conscious control or oversight with respect to some behaviour. But before we draw our attention to the similarities between these related notions, let us first enumerate several of the commonly proposed sources of automatism that have been advanced in the courts and in legal theorising. This list includes: 1- Concussion or severe physical blow to the head; 2- Hypoglycaemia[101]; 3- Somnambulism; 4- Hypnotism; 5- Psychological blow; and 6- High blood alcohol level.[102]

The sort of automatism that results from the first item on the list (i.e. 'concussion' or 'severe physical blow to the head'), is typically treated as an instance of non-insane automatism

[98] As mentioned above, the automatism defense can be advanced along two different lines; namely, insane automatism and non-insane (or sane) automatism. It should also be noted that these categories are typically treated as being mutually exclusive (see: Healy, 2000, p. 95-96).The former is counted as part of the standard defense of mental disorder, while the latter, if successful, according to Brudner, "…means that the accused has not *acted*, that his bodily movements were not expressions of a mind or will; hence it leads to an absolute acquittal" (2000, p. 67). The reason that it leads to an acquittal in such an instance is that it amounts to a lack of voluntariness on the part of the accused and that component of voluntariness is required by law for a conviction.

[99] See: http://www.justice.gc.ca/eng/dept-min/pub/md-tm/defin.html

[100] Notwithstanding, the notion of automatism and its legal application continues to be a contentious issue, as it has been for the past several decades. Indeed, its application and varied interpretation in case law has been sharply criticised along various different lines. See for example, Healy (2000); Brudner (2000); Kalant (1996).

[101] See: Holland (1982-1983, p. 112-113).

[102] See: Kalant (1996, p. 632-633).

since the effects of suffering a severe blow to the head are not normally classified as indicative of a mental disorder, but rather, they are treated as the aftermath of an external event that is not typically expected to afflict an individual in a chronic manner. Next, automatism resulting from hypoglycaemia has also been treated as being of the non-insane sort[103], since, although diabetes is a disease, it is not commonly treated as a "disease of the mind". The cause of automatism known as somnambulism (or sleep-walking in common parlance), on the other hand, may be considered to be a disease of the mind (if its occurrence is in fact caused by an underlying mental disorder), and may therefore fit best under the heading of insane automatism.[104] With respect to automatism resulting from hypnosis, it would appear that—insofar as we accept behaviours performed while under hypnosis into the category of automatism—such behaviours would fall under the non-insane heading since they too represent the consequences of a particular external event upon the individual (*viz*. the event of being hypnotized), and would thus not be expected to amount to a chronic mental condition either. When it comes to the claim of automatism being the result of a psychological blow (i.e. severe psychological shock or trauma), the issue is hotly contested. Indeed, Healy claims that: "…much controversy has surrounded the question whether a severe psychological blow could also be a cause of non-insane automatism" (2000, p. 90). Nevertheless, where a psychological blow has been accepted as a possible cause of automatism, it has not been considered to be a standing mental disorder[105]—and thus, psychological blow automatism has been categorized as being of the non-insane sort. Lastly, in cases where it has been suggested that a high blood alcohol level may result in automatism[106], it is again considered

[103] See: Holland (1982, p. 113).

[104] However, Holland (1982-1983, p. 113-114) notes that, so far, cases of somnambulism have tended to result in acquittals.

[105] And this is because the sort of psychological trauma in question is presumed to be preceded by an infrequently encountered and unexpected event.

[106] This is also a highly contested issue. For greater detail, see Dimock (2011); and Kalant (1996).

to be of the non-insane type since what is responsible for the state of automatism is brought on by an external cause (*viz*. the alcohol that is ingested) rather than an internal mental disorder.[107]

Without straying too far into the intricacies of law, it is worth noting that even though the majority of the above listed causes of automatism are reasonably treated as leading to those of the non-insane type, the default legal interpretation of the automatism defense favours treating it as a mental disorder. The reason that there exists a bias towards classifying instances of automatism as being of the insane sort is that the criteria that must be met in order to show that an instance of automatism is of the non-insane sort are exceedingly difficult if not impossible to satisfy. Indeed, according to Healy (2000, p. 97):

> At all events, it is clear that the viability of non-insane automatism will be nil unless the judge decides, as a matter of law, that the average sane person would react to the events in issue by a dissociation of mind and body as expressed in involuntary physical behaviour. The effect of this will be to eliminate the defence of non-insane automatism because it is a standard that cannot be met...To demand that the average sane person would react to the events in issue in a specified way is to preclude, by law, the possibility that this accused person actually *did* react...by a dissociation of mind and body, *even if* the average sane person may not have done so.

The complexities of how to legally classify differently produced instances of automatism aside, the above quotation mentions one of the ways in which automatism is understood in the courts—namely, as a *dissociation* of mind and body.[108] According to Kalant, "Dissociation is defined as a disruption in the usually integrated functions of consciousness, memory, identity, and perception of the environment" (1996, p. 636).[109] Moreover, he suggests that the condition of dissociation of mind from body means that consciousness is kept separate from things like the emotions,

[107] Granted, one may argue that alcoholism may be considered 'a mental disorder' but the alcohol itself must be ingested in order to produce the alleged automatism.

[108] Kalant (1996) points out that automatism is also sometimes interpreted to mean "blackout" but he argues persuasively that its being understood in this way "is a serious error of concept" (p. 640).

[109] He also points out that the category of "dissociative disorders" is listed in both the ICD-10 and the DSM-IV, and that this category includes this things like "dissociative amnesia, dissociative fugue...multiple personality disorder, depersonalisation, etc" (1996, p. 637).

behaviour, and judgement that a person normally has in response to a given situation, and that "the person is not capable of exercising conscious control over behaviour" (1996, p. 637) while in such a condition. Clearly, this characterization of the dissociation undergone by the individual during an episode of automatism bears a great resemblance to that conscious disconnect identified in the earlier outline of automaticity. Perhaps the only truly salient difference in the above construal of automatism, as Kalant puts it, is that under this understanding of dissociation the individual *lacks the capacity* to intervene and control his or her own behaviour; whereas, with respect to what has so far been said about automaticity, there has been no pronouncements about whether or not an individual always entirely lacks the ability to intervene and take control of the automatic behaviours that he or she engages in.

Now that we have considered the sorts of things that are treated as capable of leading to a state of automatism, as well as the two standard categories of automatism (insane and non-insane), along with some idea of just what the notion of automatism is commonly taken to imply (i.e. a dissociative state that renders an individual incapable of consciously controlling his or her behaviour), we may fruitfully compare and contrast the legal notion of automatism with the psychological notion of automaticity outlined earlier.

Beginning with the ways in which our understanding of automaticity may be seen to differ from the understanding of automatism, we may first notice that, whereas instances of automaticity were taken to commonly require a period of entrainment, this does not appear to be the case with respect to automatism. Indeed, I have found no mention of any necessary entrainment period for the behaviours produced by an episode of automatism in the law or legal

theorizing.[110] And this makes sense, given that several of the listed sources of automatism appear to preclude any process of entrainment.[111]

Another difference between automaticity and automatism would appear to be that, while instances of the former are not typically considered to be indicative of any mental disorder, there exists a strong presumption of mental disorder with respect to instances of the later. In other words, the kinds of automatic behaviours that people engage in under the label of automaticity are widely taken to befall the population at large and they generally fail to constitute a genuine mental disorder or chronic cognitive impairment; whereas, as was mentioned earlier, instances of automatism are standardly associated with the mental disorder provisions of section 16 of the criminal code, and there remains a strong pressure to treat even instances of non-insane automatism as symptomatic of mental disorder.

Yet another difference has to do with how each notion is interpreted or understood. For instance, when it comes to automaticity, the prevailing understanding is that conscious oversight of a given set of one's behaviours is absent but the upshot of this absence is that one's conscious attention is then liberated to focus upon other things, and this freeing up of attention can be both beneficial to the person and adaptive since it enables a person to operate in different and distinct ways simultaneously (thus broadening the scope of what an individual is capable of doing at a given time). On the other hand, automatism has been characterized as a sort of dissociation of

[110] However, Cooper (1994) reports that, in one case, an expert witness (a psychiatrist) suggested that "...a driver could be in a trance-like state induced by the repetitive stimuli experienced on long journeys on straight, featureless motorways" (p. 162). But, this same expert also maintained that such an individual would not be completely unaware of his or her surroundings. More importantly, the expert's comments did not suggest that repeatedly encountered stimuli (along with a standardly employed response) are required for automatism to occur.

[111] For instance, attempting to condition one's self to behave in a certain (criminal) way after repeatedly and voluntarily receiving a severe blow to the head is plainly absurd. In fact, the only possible exception to this that I am aware of is that professional boxer's are encouraged to 'clinch' upon being dazed from a severe blow to the head, but clinching with someone (which amounts to nothing more violent than hugging) is not normally considered criminal behaviour. Also, the automatic behaviours that are a result of somnambulism are not susceptible to entrainment (as far as I am aware). And it would appear that the same goes for psychological blow automatism since a severe psychological blow is not normally the sort of thing that can be reproduced on demand.

mind from bodily behaviour such that the individual, while in a state of automatism, is utterly incapable of controlling his or her emotions, judgements, and (most importantly) behaviours; and it is often treated as producing a state of complete unconsciousness or 'blackout' throughout its operation. Moreover, it is also typically regarded as responsible for the production of violent, dangerous, and destructive behaviours—which is, of course, why it is of concern to the courts.

Turning our attention to the ways in which the notions of automaticity and automatism are similar, we find that there exists a significant degree of conceptual overlap between the two. First, both instances of automaticity and automatism may occur involuntarily[112] (i.e. against one's desires or better judgment). Also, both of these notions imply a lack of conscious control over one's behavioural output as well as over one's assessment of the environmental factors that might serve to moderate those behaviours. Finally, both instances of automaticity and automatism may (and typically do) occur in the absence of any occurrent intention by the individual. In other words, not only is the behaviour in both cases automatic, but the initiation of such sequences of behaviour occurs automatically as well.

While the above comparison of automaticity and automatism renders it clear that these notions share certain core features (and that, therefore, work on automaticity and autonomy[113] may be relevant to the law and to legal theorizing), it has also been made clear that these two similar concepts are currently treated in very different ways by their respective disciplines or domains of application (i.e. psychology and the law), and thus, they ought to be kept separate in our understanding as we move forward in examining the concept of automaticity.

[112] However, as was previously mentioned with respect to automaticity, one may voluntarily initiate a given behavioural sequence from which one later diverts one's conscious oversight and control (thus leaving the behavioural pattern to run to completion automatically after that point).

[113] Alan Brudner has already begun analysing the relation between automatism and autonomy and how our understanding of these notions might inform legal policy. See his (2000) "insane automatism" in references.

Now that we have distinguished the notion of automaticity from the notion of automatism, in the next section, we will return to exclusively dealing with our understanding of automaticity in order to develop a more detailed and comprehensive grasp of how it has been studied and what exactly it entails.

2.3 An Overview of Automaticity

Although research on automatic mental processes dates back to the early days of experimental psychology[114], it was not until the late 1970's, following a paper by Shiffrin & Schneider entitled "Controlled and automatic human information processing" appeared in the journal *Psychological Review* that the modern boom of interest and research on automaticity began. Since the paper by Shiffrin & Schneider, each passing decade has witnessed a steady and significant growth in the number of research projects aimed at developing our understanding of automaticity and its impact on both the cognitive and social lives of human beings. According to Bargh, "Its renaissance can be traced to the introduction of a theoretical distinction between 'automatic' and 'conscious' or 'controlled' processes..." (1989, p. 3). Originally, automatic cognitive processes and behaviours were characterized as being uncontrolled (i.e. occurring outside of the purview of conscious awareness), unintentional or involuntary, and efficient (i.e. not placing a strain on an individual's presumed normally limited cognitive processing bandwidth). On the other hand, processes were considered to be conscious or controlled when they exhibited the opposing characteristics; namely, when they were intentional or voluntary, required conscious attention, and exacted a toll on active cognitive processing. Furthermore,

[114] See: James (1890); Jastrow (1906).

according to Chen, Fitzsimons, and Andersen, "Early definitions of automaticity imposed strict, all-or-none criteria for a process to be deemed automatic" (2007, p. 135). So, during the early days of research on automaticity, in principle, a behavioural phenomenon that operated outside of the conscious awareness of the individual, was efficient and uncontrolled, but was nevertheless intentional would fail to count as an instance of automaticity under this strict view.[115]

Although the mentioned characterizations of and distinction between automatic and controlled processes helped to establish the general scope of research on automaticity, it has also led to some confusion and misinterpretation among researchers. Indeed, Bargh suggests, for instance, that, "...discussing one's findings of great efficiency of a process in terms of its automaticity led others to infer (reasonably, given the all-or-none assumption) that the process also was unintentional and uncontrollable" (1994, p. 3). In other words, the rigidity with which researchers held to the initial distinction between automatic and controlled processes as always encompassing each of their respective component features was seen to obfuscate matters and led to unwarranted conclusions where specific research projects were aimed only at one or a few of the component characteristics of automaticity.

In more recent years, however, it has become clear that the initial construal of automaticity is out-of-touch with what has been revealed empirically through research. Indeed, Bargh more recently claimed that, "...mental processes at the level of complexity studied by social psychologists are not exclusively automatic or exclusively controlled, but are in fact combinations of the features of each" (1994, p. 3). And this makes sense when we consider the fact that, in order to study particular aspects of automatic behaviour and cognition, researchers

[115] Consider the notion of driving automatically. Normally, one's choice *to* drive somewhere is deliberate or intentional even if the behaviour of driving itself can take on an otherwise automatic profile at some point along the way (i.e. while engaged in the activity of driving).

have often relied upon methodologies which required that certain of the other components that were initially grouped together under the heading of automaticity be rejected. And in some studies, not only was it the case that one or more of these components were rejected, but that their opposing counterparts (originally conceived as belonging to the conscious or controlled category) were seen to play a crucial role in the phenomenon examined.[116] Additionally, it ought to be noted that while some research projects are aimed squarely at one or more component features of automaticity, they may nevertheless remain completely mute with respect to one or more of the other characteristic features.[117]

In an attempt to get beyond the overly restrictive and unrealistic views on automaticity that gripped researchers at the beginning of the renewal of interest in its study only a few decades ago, Bargh has opted for a different classificatory model. According to him, one important way in which to understand the research on automaticity is to inquire into the conditions of its occurrence. And this is because, as he says, "All automaticity is conditional; it is dependent on the occurrence of some specific set of circumstances" (Bargh, 1989, p. 7). Given that there are different background conditions that are required to produce different sorts of automatic cognition or behaviour, Bargh suggests that our classifications be set up in terms of these differences. In general, he finds that there are three primary categories under which instances of automaticity may be classified when they are considered in this way; these are: 1- preconscious; 2- postconscious; and 3- goal-dependent. Nevertheless, because our concerns extend beyond

[116] Indeed, Bargh (1994, p. 3) mentions, for instance, that even with respect to the prototypic experiments on automatic cognitive processes carried out by Stroop (1935), it was necessary that a subject direct focal attention to the target in order to produce the automatic effect (more about Stroop effects to follow in section 2.3.3).

[117] For the time being, we will bracket any concerns that this might raise for our initial characterization of automaticity provided at the outset of this chapter (in section 2.1). Later on, in section 2.3.3, when we are equipped with a more detailed understanding of the types of research on the different aspects of automaticity, we will see that framing our understanding of automaticity in different ways (as may be required to answer different empirical questions) does not pose a threat to our initial construal since, for example, though the element of efficiency might be studied in isolation, there nevertheless remain many examples wherein the majority of the features of automaticity identified in section 2.1 are required to produce the behaviours in question.

merely identifying and understanding the requisite conditions for the production of various sorts of automaticity, for our purposes, we may narrow down these categories to only two. First, we may treat both Bargh's preconscious and postconscious categories as falling under what I will call the 'percept-judgement' category; and second, we may treat Bargh's goal-dependent category as falling under our 'behavioural' category.[118]

2.3.1 Percept-Judgment Automaticity

Beginning with what Bargh labels the preconscious sort of automaticity, he claims, "A preconscious automatic process requires only that the person notice the presence of the triggering stimulus in the environment" (1994, p. 4). And this 'noticing' of the triggering stimulus can happen either at the level of a conscious perception or at the level of a subconscious (sometimes called 'subliminal'[119]) sensory input. Moreover, the processes that are initiated in this way can happen independently of any goal or intention on the part of the individual. Indeed, according to Chen *et al.*, "…these processes occur immediately upon registering the stimulus, and they are completed before perceivers grasp, if they ever do, that such a process has occurred" (2007, p. 135). So, with respect to the early core component features of automaticity mentioned above, preconscious automatic processes are in accord with the requirements of being efficient and unintentional or involuntary, but they remain ambiguous with respect to whether or not the

[118] While it is true that all instances of automaticity are evinced only insofar as they are observable in behaviour, depending upon the sort of automatic functioning studied, the research has typically placed greater emphasis upon either the perceptual and judgmental component or that of behaviour. In any event, what is most pertinent to our investigation is automatic behaviour, and so, that is what we will be primarily concerned with in the following.

[119] However, Bargh warns that, "...preconscious is not synonymous with subliminal, although subliminal processes are certainly a subset of preconscious ones" (1994, p. 4).

individual is consciously aware of the triggering stimulus—in some cases the individual may be consciously aware of the stimulus, in other cases conscious awareness is not needed for the effect to occur. Nevertheless, in either case (i.e. conscious of the stimulus or not), the individual typically does not control the cognitive or behavioural effect of the registered stimulus.

What remains to be said, however, is just what types of processes are begun in this sort of direct and entirely non-deliberative manner. According to Bargh, the list of automatic cognitive processes that may be initiated preconsciously includes "interpretations, evaluations, and categorizations" (1994, p. 4)—which, for our purposes, may be generally understood as kinds of judgments (hence, these processes are reasonably subsumed under the 'percept-judgment' label). Examples of these processes include studies on attitude activation, attention responses to negative stimuli, physiological reactions, and frequently available trait construct effects upon social perception.[120]

The next category that Bargh identifies is what he calls postconscious automaticity. He suggests that the sorts of effects that are produced by instances of postconscious automaticity mirror those that are produced by preconscious automaticity, and that the main feature that distinguishes the one from the other is that postconscious automaticity depends upon recent conscious attentional processing whereas preconscious automaticity does not (1994, p. 5)[121]. Despite this difference, Bargh claims "Postconscious effects are functionally the same as preconscious ones, except that they are temporary…" (1994, p. 5). In other words, both of these sources are capable of initiating similarly automatic cognitive processes, however, in one instance (*viz.* the postconscious one) the effect is a result of an individual having activated a given conscious cognitive process in close temporal proximity to the observed effect, and its

[120] See Bargh, (1994, p. 4-5; 1989, p. 11-14) for a list of the relevant studies.

[121] Studies of how priming effects impact impression formation stand as prototype examples of postconscious automaticity. We will consider an exemplar of these sorts of studies later in this chapter (in section 2.3.3).

resulting automatic influence upon an individual's subsequent judgments tends to be of a transitory or fleeting nature.

2.3.2 Behavioural Automaticity

The last category that Bargh distinguishes is goal-dependent automaticity. He claims that this "…class of automatic phenomena only occurs with the person's consent and intent" (1994, p. 6).[122] To illustrate, Bargh (1992) provides the example of a character named Otto. In short, Otto is portrayed as driving around in a new town and upon rounding an abrupt turn, he catches a glimpse of a stop sign and his foot automatically stomps on the brake pedal just in time to save him from traversing into the oncoming traffic of a busy intersection. At a later date, Otto goes for a walk along this very same roadway and rounds the same corner. This time, however, Otto's leg does not kick out automatically upon registering the stop sign since his behaviour is not constrained by the same operational goal (i.e. this time he is not driving and thus kicking out as if to apply a pedal brake would be entirely useless and incoherent). In the first place (i.e. when Otto was driving), his automatically stomping on the brake can be construed as being in accord with or following from his operational goal since he intended to safely drive around in a new locale in order to familiarize himself with his new surroundings. And it seems clear that this over-arching intentional goal was in some way responsible for the specific and appropriate automatic

[122] However, in a 1989 article (see references), he does distinguish between unintended and intended goal-dependent automaticity, wherein he claims that, "Unintended goal-dependent automatic effects have as a necessary precondition the instantiation of specific processing contexts, but they are unintended consequences of those intentional thought processes" (1989, p. 20). In other words, the requisite 'intent' is not directly concerned with the automatically actuated behaviour; rather, it has to do with the background operational goal that makes sense of the automatic behaviour. Expressed differently, the agent does not possess the occurrent conscious intention to bring about the automatic behaviour that is produced but only the over-arching behavioural aim which then shapes and constrains the automatic behaviour that takes place.

behavioural response (*viz*. immediately stomping on the brake pedal) to the abruptly encountered stimulus. Experimental examples of automatic behaviour that fall under the goal-dependent heading for Bargh are drawn from research on things like behaviour-to-trait judgments, self and other trait concepts, implicit learning, incubational processing, and action slips.[123]

Though it is clear that some of the experimental examples that Bargh provides of goal-dependent automaticity continue to be more emphatically related to forms of judgment and other cognitive processes, it is also the case that this category of automaticity is most closely connected to our concern with personal and autonomous agency—since, the category includes examples concerned not only with cognitive processing goals but with behavioural goals as well (and this is why I have opted to use the label 'behavioural automaticity' for such cases). This category is also the one that lines up most directly with our earlier construal of automaticity (and its status as a potential threat to autonomy) provided at the outset of this chapter. Thus, we may consider the kind and the characteristics of the sort of automaticity described in section 2.1 as falling under the heading of behavioural automaticity and fleshing out this category.

In order to increase our understanding of the different ways in which automaticity has been studied, and to connect the distinctions made above with some of the empirical research that has been done, in the next section, we will consider some of the key types of studies aimed at elucidating various aspects of automaticity.

[123] See Bargh, (1994, p. 6; 1989, p, 19-28). The research on action slips will be examined in greater detail in section 2.4.

2.3.3 Examples of Automaticity in the Lab

One of the most well known series of experiments to have revealed what Bargh terms preconscious automaticity is the study executed by J. R. Stroop in the mid-nineteen thirties.[124] Stroop's research took its lead from earlier work on interference effects upon habitual cognitive processes as well as research focused upon differential processing and verbal report times for identifying colour stimuli *vs*. colour names.[125] In essence, Stroop's project combined what was gleaned from this earlier research into a single study (consisting of three related experiments) that would seek to establish what sorts of interference effects might arise in verbal report tasks that consisted of non-paired colour names and ink colours. In the first experiment, participants were presented with a list of colour words that were printed in an ink colour that did not match the target words (for instance, the word 'blue' may be written in red ink and *vice versa*, the word 'green' might be written in purple ink, *et cetera*). The participants were then asked to read the colour word as quickly as they could (and the total time that it took them to do so was contrasted with the time it took for them to read the same list printed entirely in black ink). In a second trial, participants were asked to state the colour of the ink as quickly as they could despite the colour of the ink being different from the colour identified by the list word (and this was compared with the participants' time in reporting the colours of a similar list that consisted of solid coloured squares instead of letters).[126]

[124] Indeed, the sorts of effects identified in his findings have since come to be known as "Stroop effects".
[125] See Stroop (1935, p. 643-647).
[126] The third experiment in this series aimed to determine whether or not any noted effects would change due to practice. See Stroop (1935), for greater detail with respect to the experimental protocol and findings.

According to Stroop (1935):

> The increase in time for reacting to words caused by the presence of conflicting color stimuli is taken as the measure of the interference of color stimuli upon reading words…[whereas]… (t)he increase in time for reacting to colors caused by the presence of conflicting word stimuli is taken as a measure of the interference of word stimuli upon naming colours. (p. 647)

This experimental design enabled Stroop to compare the interference effects of each sort of stimuli upon the other while using the same list pairs of stimuli across experiments. Interestingly, the interference produced by the differing colour stimuli were not found to be reliable (adding only 2.3 seconds to the reading time for a hundred colour word list). However, it took participants on average 47 seconds longer to correctly report the printed ink colours of the one hundred word list when those words identified conflicting colours. Because the interference caused by the colour stimuli was negligible while the interference produced by the word stimuli was highly noteworthy, Stroop concluded that, "…the associations that have been formed between the word stimuli and the reading response are evidently more effective than those that have been formed between the colour stimuli and the naming response" (1935, p. 659-660). What this means with respect to automatic cognitive processes is that the mere presence of a conflicting word stimulus was seen to automatically interfere with an individual's response time in identifying the ink colour of the printed word. These findings fit into Bargh's category of preconscious automaticity since what is required is only that a person notice the environmental stimulus—in this case a word with a conflicting colour designation from the colour of the ink in

which it is printed—in order to produce a given cognitive/behavioural effect (in this case, a cognitive interference that led to delayed verbal report).[127]

Another sort of experiment commonly used to study automatic processes has to do with what are known as 'priming effects'. Studies that rely on priming methodology produce effects that fall under Bargh's category of post-conscious automaticity mentioned earlier—and this is because they require "…conscious experience or thought in the same stimulus domain as the automatic process…" (Bargh, 1992, p. 190) prior to the initiation of said automatic process. In these sorts of experiments, participants are surreptitiously 'primed' by exposure to information aimed at activating a given cognitive representation in one task prior to engaging in another task that they are instructed is unrelated to the first but that is, unbeknownst to the participants, designed to elicit measurable automatic processes that are related to the primed information. What typically follows is that the primed participants show a significant bias in favour of the primed representation in the later task versus a control group. For instance, in a study by Skelton and Strohmetz (1990), participants in an experimental group were primed with a word task aimed at identifying health related words prior to being given the Pennebaker Inventory of Limbic Languidness (PILL), a 54 item checklist for a variety of physical symptoms. The results of this study confirmed the hypothesis that participants primed with the health related word activation reported a much higher number of symptoms on the PILL questionnaire then subjects in the control group.[128] In another study, Mark, Sinclair, and Wellens (1991) revealed that participants who were first given the Beck Depression Inventory (BDI)—a self-report measure

[127] It is important to note here, as Bargh does, that "…the subject does not intend and cannot control the interference caused by the meaning of the stimulus word…[yet]…(t)his [automatic] interference effect…does not occur without the devotion of spatial attention to the word's location" (1992, p183). And because it requires a certain degree of focused attention in order for the effect to be produced, the Stroop effect stands as an example of an automated cognitive process that does not fit into the original and rigid "all-or-nothing" definition of what is required for automaticity.

[128] See the Skelton and Strohmetz (1990) study listed in references for greater detail.

of depression—later reported either a more negative mood assessment (if they were already part of the more depressed group) or alternatively a more positive mood assessment (if they were part of the group of nondepressed participants), when contrasted with similarly divided control groups.[129] Again, what these results show is that merely being primed (in this case by the negative content of the BDI) was sufficient to significantly affect a participant's mood-state self-assessment when compared to controls. Moreover, because the participants in these studies were kept ignorant about the experimental aims, it is presumed that participants were not aware of the automatic influence of the primed information upon their later judgments.

The last type of experiment that I will mention in this section falls under Bargh's category of goal-dependent automaticity. What sets this category apart from the previous two for Bargh and Chartrand is that "much, if not most, of our responses to the environment in the form of judgments, decisions and behaviour are determined not solely by the information available in that environment but rather by how it relates to whatever goal we are pursuing" (1999, p. 468). What this means is that if we are in the process of trying to reach some end or desired state of affairs, our automatic behaviours and evaluations may continue to be responsive to environmental triggers and changes but they will also be constrained by what is prescribed by those very objectives. In this sense, as with the post-conscious category, a certain degree of intentionality or conscious processing is required in order to produce the automatic behaviour.[130]

[129] See the Mark, Sinclair, and Wellens (1991) study listed in references.

[130] But this view simplifies things somewhat. Goal-directed forms of automaticity are not all the same; some are intentional in the sense that one must first consciously initiate the pursuit of some end before consciousness can withdraw thus allowing automatic processes to take over and see the selected behavioural goal through to its completion, whereas others may be constituted by habitual behaviour patterns that are straightforwardly provoked by some familiar environmental feature. And, insofar as those habitual behaviours were at some previous point in time consciously formed, one might still opt to call them 'intentional' but there is a clear difference between these two forms of automaticity. See Bargh (1989, p. 19-29) for more information about the various forms of goal-directed automaticity.

In a relatively straightforward experiment, Aarts and Dijksterhuis (2000) set out to determine whether the activation of a habitual travel destination goal would also activate an automatic travel mode selection. They maintain that, "…habits are automatic goal-directed behaviours that are mentally represented…as strong associations between goals (e.g. going to the supermarket) and actions (e.g. using a bike)" (2000, p.76).[131] Thus, they hypothesized that the activation of a habitual travel destination goal would in fact automatically prompt a habitual travel mode selection (be it to walk, or take a bike, train, or bus) and that, consistent with prior research on automaticity, it would be difficult to suppress the automated habitual travel mode selection. They began by conducting a pilot study of thirty Dutch university students living in the city center of Nijmegen. These students were given a list of 50 travel destinations and asked which mode of transportation they would usually take to get to each of the destinations. This provided data about the top five routes these students most commonly and frequently travelled by bicycle (e.g. a student dance club called the Swing, the university, the sports center, *et cetera*) and which, with respect to their selected travel mode, would presumably be the most difficult to suppress. They also determined the top five locations that were external to the city and for which students would typically take the train but that remained destinations that students would not frequently visit (thus these locations represented nonhabitual destinations).

With the above mentioned data from the pilot study in hand, Aarts and Dijksterhuis performed a study with fifty-six undergraduate students from the University of Nijmegen as participants. Their study followed a 2x2x2 design: (typical travel mode: permitted vs. not

[131] It is important to note along with Wood and Neal, however, that "Habits typically are the residue of past goal pursuit; they arise when people repeatedly use a particular behavioral means in particular contexts to pursue their goals. However, once acquired, habits are performed without mediation of a goal to achieve a particular outcome or a goal to respond" (2007, p. 844). In other words, although goals may shape habits in the early stages of their formation, once a certain sort of behaviour becomes habitual, it no longer requires the presence of an occurrent intentional goal to become active. To put it differently, Wood and Neal state that "...habits may be goal directed [i.e. structured by earlier goals]...even though they are not goal dependent [or contingent upon the occurrent activation of a goal]" (2007, p. 847). More on habits later in section 2.5.

permitted); (cognitive load: yes vs. no); and (habit: yes (bicycle) vs. no (train)) (2000, p. 77). The participants would take part in the study by computer and were separated into different cubicles. They were told that the study was about travel behaviour and they would be questioned about their typical travel modes in relation to a list of destinations both inside the city center and outside of it. One of the experimental manipulations involved had to do with the typical mode of travel: half of the participants were instructed to state which mode of travel they would typically use to arrive at the given destination (this was the typical travel mode permitted group), while the other half were instructed not to state their typical mode of transport but instead to select another mode (this was the typical travel mode not permitted group). The next experimental manipulation had to do with cognitive load. Half of the participants were given the task in the context of a secondary and cognitive resource depleting prior task. This secondary task involved summing two digits that would appear at either end of the travel destination word prior to mentioning the travel mode (e.g. 6 sports center 9). These participants were part of the 'cognitive load' condition whereas those not burdened by this additional task were part of the 'no cognitive load' condition. Moreover, all of the participants had just three seconds to respond to each trial destination word and, if they failed to answer within the allotted time frame, a tone would sound and the next target word would appear on the screen.

In short, what Aarts and Dijksterhuis found was that where participants were not permitted to respond with the typical mode of transport and were also under cognitive load, they provided the bicycle response (to the previously established top five bicycle destination pairs) at a significantly higher rate than those participants that were not permitted to respond with the typical mode of transport but that were not under the cognitive load condition.[132] According to Aarts and Dijksterhuis, "In general, these failures to suppress corroborate the notion that the

[132] See Aarts and Dijksterhuis (2000) for more detail.

bicycle responses are automatically triggered by travel goals" (2000, p. 79). In other words, the automatic selection of the habitual travel mode 'bicycle' to commonly and frequently paired travel destinations appear to be constrained by the activation of the destination goal. These travel mode selections not only appear to be constrained by the habitual pairing of travel destination and travel mode but they also appear to be automatically activated by the given travel destination goal. Thus, the findings mentioned above exemplify Bargh's goal-dependent category of automaticity.

The experiments outlined above are by no means a complete representation of the empirical research on automaticity. Indeed, they represent a mere snapshot of some of the work that is available on the subject. However, each selected experiment, I think, reveals something of the different ways in which an instance of automatic thought or behaviour can be brought about and expressed. Moreover, while I have tried to follow Bargh's classificatory lead in order to make clear the defining features of various forms of automaticity, it remains the case that, depending upon one's experimental objectives, one may characterize automaticity in a number of different ways—so long, that is, as it concerns at least some feature of the originally outlined construct. That is to say, depending upon one's research goals, one may operationally define automaticity along a number of different lines. Additionally, there is the problem that, in some cases, a preconscious situational trigger may be what sets off an instance of goal-dependent (or behavioural) automaticity. So there is a sense in which Bargh's (and my own) categories may collapse; that is to say, an instance of automaticity may be constituted by central features from more than one of the proposed categories. In other words, there may exist hybrid forms of automatic behaviour—that are, for instance, both preconscious and goal-dependent—out in the world, even if examples of them have not yet been documented in the lab. Nevertheless, Bargh's

three mentioned categories, as well as my own two more general classifications, do enable us to recognize some of the central and defining features of some of the primary kinds of automaticity, and they also serve as a worthwhile framework from which to approach the experimental literature in order to differentiate between the experimental aims of researchers. Moreover, an appreciation of what is characteristic to those examples that fall under the category of behavioural automaticity allows us to sharpen our focus upon matters that are directly relevant to our concerns with personal autonomy.

As was previously mentioned, I am primarily concerned with goal-dependent or goal-directed forms of automaticity as these are expressed in behaviour but this does not mean that instances of automaticity that fall under what I call the percept-judgment category are of no concern to a theory of autonomy. Rather, I think that even automatic percept-judgments are a threat to personal autonomy. However, I maintain that the sorts of automatic processes that are initiated in that kind of way alone (i.e. entirely outside of the person's awareness of their causal influence upon behaviour) fall under our procedural independence condition outlined in section 1.2.2, and thus are unambiguously to be stricken from counting in any way as expressions of a person's autonomy. What concerns me instead are instances of behavioural automaticity that do appear to embody some degree of intentional structuring (even if the intentionality in question is temporally remote from the issuance of the behaviour) since these cases are not so obviously handled by the procedural independence condition.

One aspect of behavioural automaticity that helps to shed some light on just why we might not want to treat even goal-dependent forms of automaticity as amounting to genuine expressions of autonomy can be found in what are referred to as 'action slips' or 'actions not as planned' in the literature. In the next section, I will explain just what action slips are, as well as list some of

the various ways in which they can manifest. Later, I will provide some examples where both action slips and other forms of behavioural automaticity would reasonably lead us to conclude that automatized behaviours should not count as expressions of an individual's autonomy.

2.4 Actions Not as Planned

By now the notion of a 'slip of the tongue' (often referred to as a 'Freudian slip') is more or less commonplace. The phrase refers to the accidental, erroneous, or absent-minded vocalisation of a word or statement that a speaker intended to keep private or at least did not intend to say out loud, and which frequently results in the speaker feeling embarrassed or ashamed. Freud thought that these slips of the tongue betrayed some unconscious need or desire on the part of the speaker but a simpler account might treat them as merely the innocuous or at least unintended results of strong—i.e. frequently and recently activated—thought associations or habitual elements of mind that have come to occur together automatically.[133] As common as the notion of a slip of the tongue is, there is a related behavioural notion that would appear to be much less widely recognized. Indeed, the notion of a slip of action, it seems, is far less widespread than the notion of a slip of the tongue; and this would appear to be the case despite the fact that people often claim that they "didn't mean" to act as they did (following an action

[133] Recall the experiment conducted by Aarts and Dijksterhuis (2000), wherein participants under cognitive load were often unable to give a response other than 'bicycle' as the mode of transport to five locations most commonly and frequently travelled to by bicycle even when they were instructed not to respond with their most typical mode of transportation to these locations. These participants may not have conceived of their responses as slips of the tongue, but they nevertheless appeared to automatically draw upon a habitual association between destination and mode of transport when under the added pressure of cognitive load. Likewise, people often experience slips of the tongue when under the nervous pressure to impress another person and the mindful attention to what not to say, it would seem, could plausibly act as a priming like cognitive rehearsal that may end up strengthening whatever associations in thought and speech one is trying to avoid rather than keeping one from saying whatever it is that one wishes to refrain from saying.

slip) similarly to how, following a slip of the tongue, they might claim that they "didn't mean" to say what they had just said.

Given the similarity of the typical types of reactions produced by these two phenomena (i.e. once the speaker or actor becomes cognizant of the unintentional slip) we might reasonably consider each sort of occurrence to be an expression of the more general phenomenon of automaticity. In addition to the shared sorts of reactions that such occurrences evoke, each notion is also similarly defined. Indeed, action slips may be characterized as referring to the accidental, erroneous, or absent-minded behavioural sequences that an individual performs in the absence of the occurrent intention to do so. And, just as with slips of the tongue, an action slip can leave an individual feeling embarrassed, ashamed, or simply regretting that the action slip occurred.

In an important 1979 paper entitled: "Actions Not as Planned: The Price of Automatization" which aimed at elucidating the sorts of behaviours that may fall under the heading of 'slips of action', the psychologist James Reason contrasted actions not as planned with actions that in fact do successfully carry out a plan. Reason recognized that although an individual may sometimes fail to reach a particular goal that he or she has selected, it is not always the case that such failures are the result of an error to behave according to a plan on the part of the individual. Indeed, sometimes, unforeseen or unforeseeable interferences may thwart one's attainment to a given end, and in these cases the interfering elements may be seen as responsible for the individual's missing the mark rather than some behavioural shortcoming on the part of the individual. On the other hand, not all actions that result in the attainment of a given goal ought to be considered non-errors either. For instance, as Reason suggests that, "...in the case of the golfer whose misdirected ball is deflected into the hole by a passing bird" (1979, p. 152), the successful achievement of the end aimed at is more accurately described as a matter

of luck, or accident, than a successfully carried-out plan—since, a bird's being in just the right spot at just the right time to deflect an otherwise wayward ball is not the sort of thing that one can confidently include in one's planning. When it comes to how one might fail to meet a goal, Reason maintains that this can happen in one of two ways: first, when one behaves according to a plan but the plan was insufficient to reach the goal; and second, when the plan is sufficient but one's behaviour fails to conform to the plan. It is the second of these ways of failing to reach a goal that is central to the notion of action slips or actions not as planned.

In order to determine just how frequently action slips occur, and to obtain a broader understanding of the ways in which individuals commit action slips, Reason performed a two week long diary study whereby participants were instructed to provide a continuous daily record over the course of the study that would chronicle their "...unintended or absent-minded actions, and...[their] 'accidental behaviour' in general" (1979, p. 153).[134] Thirty five participants took part in this study (twenty three women and twelve men), and the average was just over twelve action slips per participant over the course of study. The sheer variety of the reported instances led Reason to develop a number of classifications in order to begin to identify some of the common features of the reported incidents in the hopes of isolating their causes; and it is to these classifications and their features that we turn to next.

To begin with, Reason grouped the reported incidents under several general headings, the first of which being "discrimination failures" (concerning objects in one's environment). These included *perceptual confusions* resulting from object similarity (e.g. putting shaving cream on one's toothbrush instead of tooth paste because both tubes have a similar appearance); next were *functional confusions* ostensibly due to objects sharing a similar function (e.g. a sunbathing participant intended to head outdoors with a pair of sunglasses and to leave the suntan lotion

[134] See the article listed in references for greater detail.

indoors but instead left the sunglasses indoors and took the lotion outside); followed by *spatial confusions* that resulted from objects being close to one another (e.g. grabbing a fork from the utensil drawer to butter bread instead of the intended butter knife); and finally, *temporal confusions* which resulted in unsuitable actions traceable to misperceptions of time (e.g. putting on one's work clothes upon getting up and heading out the door before realizing that it is Sunday and that one does not work on Sundays).

The next general category was "program assembly failures" (which concern the operation of behavioural programs). The first item listed under this category was what Reason called *behavioural spoonerisms*—which were essentially reversals of behavioural programs (e.g. attempting to drink from a bottle before taking the lid off)[135]; next were *active program confusions* (e.g. putting the dirty dishes in the fridge and the leftovers in the sink after preparing a meal); another was *confusions between current and stored programs* (e.g. looking for a calculator when one's new phone has one built in and is readily accessible).

Another important category identified by Reason was what he called "test failures". Test failures had to do with errors that occurred at various 'checkpoints' in the progress of a given behaviour sequence. The first of these was labeled *stop-rule overshoots* wherein a participant would continue past some initial intended end point (e.g. going out to pick-up the news paper to read with one's morning coffee but then proceeding to put it immediately in the recycling bin once inside); next was *stop-rule undershoots* wherein an individual stops a behaviour sequence prior to reaching the aimed at end point (e.g. stepping into the bath prior to entirely disrobing); followed by *branching errors* wherein multiple different goals are begun by the same early

[135] In one of the more humorous diary entries in this study, a participant reported that: "When I leave for work in the morning I am in the habit of throwing two dog biscuits to my pet corgi and then putting on my earrings at the hall mirror. One morning I threw the earrings to the dog and found myself trying to attach a dog biscuit to my ear" (1979, p. 155).

behavioural sequence but one's behaviour diverts somewhere along the way to an inappropriate alternate goal (e.g. intending to drive to one location to find oneself in route to another); and finally, *multiple side-tracking* whereby one deviates from one's goal and undertakes various other minor tasks.

The final two major categories listed by Reason are "sub-routine failures" and "storage failures". Sub-routine failures include *insertions* (i.e. adding unwanted or unnecessary behaviours to a sequence); *omissions* (i.e. leaving requisite actions out of a behaviour sequence); and *misordering* (where one performs the appropriate behaviours to accomplish the goal but does not perform them in the correct order). Storage failures, on the other hand, include *forgetting previous behaviours* (this can result in one losing one's place in a behavioural sequence); *forgetting discrete items* (e.g. planning to pick-up bread, milk, eggs and butter from the grocery store but heading home without the butter); *reverting to earlier plans* (e.g. intending to go rent a movie before remembering that one's movie player is broken and then heading out to pick-up a movie anyhow); and last, *forgetting the substance of the plan* (e.g. heading into the kitchen and forgetting what it was that brought you there).

As evidenced by Reason's study, there are clearly many ways in which one's behaviour can fail to follow a plan. But the mere compiling and categorization of such accounts does not yet tell us why these sorts of action slips tend to occur. To answer that question, Reason provides three notions that he thinks help to explain how it is that action slips come about. The first notion has to do with the 'mode of control' of a given set of behaviours. He argues that as we learn and develop various motor skills we initially rely upon what he refers to as a "closed-loop" mode of control—a mode of control that is characterized by a heavy reliance upon visual and proprioceptive feedback, conscious attention, and which places a significant demand upon one's

apparently limited conscious processing capacity in order to control one's moment to moment behavioural output. Whereas, in the "open-loop" mode of control, Reason claims that, "…motor output is governed by "motor programs" or pre-arranged instruction sequences, that run off independently of feedback information" (1979, p. 158) and which free-up central (i.e. conscious) processing.[136] It is important to note here that, similarly to Bargh's current view on the potential for a more complex interplay of different uncontrolled and controlled elements in the production of various forms of automaticity, Reason maintains that it is most likely the case that any given series of behaviours, skilled or not yet skilled, will involve elements of both closed-loop and open-loop modes of control at different times. Thus, Reason's view appears to be in line with the more recent thought on automaticity rather than the sort of all-or-nothing traditional view.

The next important component to play a role in explaining action slips for Reason is the notion of 'critical decision points'. One thing that is essential to acting according to a plan, for Reason, is that at various key points in a planned series of behaviours the individual is able to successfully resort to the closed-loop mode of control when required. Some of the examples of the 'test failure' kinds of action slips help to make this clear—for instance, the individual who stepped into the tub before entirely disrobing would have benefitted by closed-loop control prior to stepping into the tub since, this sort of control could have made apparent the need to finish undressing first. Likewise, the individual who went outside to pick up the morning paper to read with breakfast would have been well served by closed-loop control prior to tossing the paper into the recycle bin. Again, in this case, the closed-loop mode of control could have rendered it apparent that the paper had not yet been read. The problem here is that when certain types of

[136] Reason's notion of the different 'modes of control' should sound familiar since they more or less reiterate what was said in section 2.1 about the role of automaticity in relation to skill acquisition and practice.

action slips occur, they appear to be the result of a failure to initiate closed-loop control at certain points where that form of control is essential to successfully carrying out a plan.

The final important feature to recognize, according to Reason, is the 'strength' of the motor program. According to Reason "One of the most consistent findings of the diary study was that when actions deviate from current intentions they tend not to take the form of isolated and novel fragments of behaviour, but of well-practiced and functionally intact behavioural sequences" (1979, p. 160). So what appears to be happening during certain sorts of action slips, suggests Reason, is that, if a closed-loop mode of control is not successfully activated at one of the critical decision points, then some sort of conditioned motor program takes over and keeps the individual actively behaving, even though the behaviour that such a motor program supports does not accord with the more recent goals or plans of the individual. Moreover, it appears to be the case that it is the frequency and recency with which a given motor program has been activated that determines which behavioural sequence obtains during an action slip.[137] Also worth noting is that this ostensible motor program take-over of behaviour is most prominent, according to Reason, when "the central processor is occupied with some parallel mental activity" (1979, p. 160) as we saw was the case in the Aarts and Dijksterhuis (2000) study.

With the exception of the identification of the critical decision points that appear to require a closed-loop mode of control (i.e. conscious awareness and intervention) in order to sustain a series of behaviours initially aimed at achieving a particular goal, nothing that Reason says about the features of action slips should sound unfamiliar. And this is because action slips are primary examples of behavioural automaticity, so it makes sense that what has the potential to lead to action slips is consistent with what has so far been said about the production of behavioural

[137] Again, the importance of the frequency and recency of a given set of behaviours to the production of action slips should sound familiar since both of these elements have already been identified as central determinants of automated behaviours.

automaticity in general (i.e. in terms of the lack of conscious control and the frequent and recent pairing of event type or goal and standard behavioural response). But before going on to elaborate as to why it is that I think that we ought to not treat instances of behavioural automaticity as even remotely representative of an individual's autonomy, I would first like to say something about the component feature of automaticity that is characterized by the frequent, consistent, and recent pairing of an event type or goal and a common behavioural response and how these features relate to what are generally thought of as habits.

2.5 Habit

Quite clearly, what has been said so far in this chapter about what is involved in learning a new skill or conditioning one's self to behave in a certain way by frequently, consistently, and recently performing a particular behavioural response to a given set of circumstances or performance goals is deeply consistent with how habits are formed. Indeed, the similarities between what goes in to producing both automatic and habitual responses can lead researchers to conflate the two notions or assume that because a particular behavioural response is habitual that it must also be automatic. For instance, in their study mentioned in section 2.3.3, Aarts and Dijksterhuis make this assumption when they provide their operational definition of habit (2000, p.76). However, their experimental design and findings tend to support the notion that the habits that they were examining were in fact automatic (because of the degree to which participants failed to suppress the habitual responses).[138] Nevertheless, one ought to be careful when making such assumptions since a particular set of behaviours might conform to a habitual set and yet not

[138] The failure to suppress these responses suggests that they were uncontrolled or uncontrollable.

be automatic. For example, one might have a habit of going for a morning run along a particular pathway but over time one may grow bored enough with the routine to actively consider a different trajectory on a given morning. Upon reflection, one might nevertheless decide that one's habitual route remains either the safest, or the most pleasant, or in some other way the more preferable route in contrast to the other perceived route options; and thus, one might consciously decide to stick to the routine despite one's being bored with that particular course. In such a case, the morning run behaviour is consistent with a prior habit but it is nevertheless not automatic since the individual consciously selected to behave in conformity with that habit. Nevertheless, the possibility remains that on some later day the individual may fail to consciously consider which route to run and unreflectively allow an unconsidered habit to automatically dictate which course will be taken.

In addition to the danger of conflating the notion of habit with the notion of automaticity, there is another familiar issue to keep in mind. Like the notion of automaticity (as well as that of autonomy), the notion of habit has meant different things to different researchers. And like the concept of automaticity, one may operationally define habit in a number of different ways.[139] Indeed, according to a study by Clark *et al.* (2007), there are at least nine dimensions (some with a certain degree of overlap) along which habit has been studied; these include the following: 1- habit as tic; 2- habit as conditioned responses; 3- habit as addiction; 4- habit as routine; 5- habit as character, *et cetera* (p. 9S). Therefore, again, when examining the notion of habit in relation to the notion of automaticity, one ought to remain careful and clear about the particular understanding of habit with which one is working. For our purposes, the "habit as conditioned

[139] See Clark, Sanders, Carlson, Blanche, & Jackson (2007) for a summary of some of the many ways in which the concept of habit has been used by researchers.

responses"[140] and "habit as routine"[141] categories described by Clark *et al.* are the most relevant to the development and understanding of automaticity with which we are dealing. However, the general features of being "...relatively unconscious, nonreflective, and repeatable" (2007, p. 15S) identified by Clark *et al.* as being of the "core essence of habit" are also important to us.

Analogously to what has already been said about automatic behaviours in general, Clark *et al.* see habits as being advantageous to an organism since habits enable an individual to "...conserve energy by circumventing the need to allocate resources continually when the organism confronts similar situations over the course of time" (2007, p. 15S). Other benefits of habit according to Clark *et al.* include the elimination of the need for decision making and a greater speed of action or response time. However, they do recognize that habits have a negative aspect as well which is evidenced by things like alcohol and drug addiction that are born of habitual substance use and that end in either or both physiological and psychological dependency (p. 18S). Nevertheless, Clark *et al.* maintain that habits are overall more beneficial and adaptive than they are harmful.

Although the notion of habit has been characterized in a number of different and not always entirely compatible ways, we may nevertheless attempt to provide a definition of habit that captures as best we can those elements that speak to our understanding of and concern with automaticity. The following definition provided by Wood and Neal (2007) is well suited to such an end:

[140] This view of habit is derived from the early behaviourist model which relies upon the repetitive pairing of stimulus and response.

[141] This view of habit focuses on an individual's complex behavioural activity in the world. According to Clark *et al.*, "Routines involve ordering, sequencing, and combining several simple activities to create order in one's life" (2007, p 12S). In other words, habitual routines concern a larger series of interrelated activities than do basic one off habitual behaviours.

> Habits are learned dispositions to repeat past responses. They are triggered by features of the context that have covaried frequently with past performance, including performance locations, preceding actions in a sequence, and particular people. (p. 843)

To this list of contextual features I would add that it is not only context locations that may trigger habitual (and automatic) behaviours but even single or multiple separate objects within a given environment—or one or more aspects of a given situation—that may do so. Also, it is not just the presence of particular people that might engender a habitual behaviour, but even simple statements or suggestions (made by someone present or communicated electronically or by other means) which may produce a habitual response (as was evidenced in the Aarts and Dijksterhuis study described in section 2.3.3). One's own ostensibly unprovoked thoughts might also indirectly lead to the initiation of some habitual behaviour—and by 'indirectly' I mean that the thought might be responsible for causing the habitual (and automatic) behaviour in the absence of any conscious intention to behave in the resultant habitual manner. To borrow from James, for instance, one may think about changing into one's dinner clothes upon returning home from work and unintentionally end up climbing into bed after disrobing since to do so is consistent with the prior habit of getting into bed upon disrobing in the evening (1890, p. 139). In this sort of case, the resultant behaviour of getting into bed appears to be both automatic (i.e. since it is unintentional and involuntary) and in conformity with an existing habit.

Habitual and automatic behaviours share another deep similarity which has to do with the relationship that each has with goals and their attainment. Indeed, as Wood and Neal (2007) recognize, "Habits are the residue of past goal pursuit; they arise when people repeatedly use a particular behavioural means in particular contexts to pursue their goals" (p. 844). However, they argue that once habits are formed, they no longer require the oversight of consciously adopted goals in order to produce the behaviour needed to reach a particular end. And this mirrors what

has already been said about the sort of withdrawal of conscious attention which leads to automaticity and that can follow upon having learned a new skill through repetitious practice. To learn a new skill or to develop a new habit one must, at the outset, often consciously be concerned with achieving a particular goal; this goal may then lead to a frequently repeated behavioural response selection (i.e. with respect to developing a habit), or to effortfully practicing at performing specific behaviours in specific circumstances or at specific times or places (i.e. with respect to developing a skill). But, upon achieving a certain degree of competence, fluency, or conditioning, one need not rely upon those early goals to produce the specific behaviours any further. However, despite the eventual unnecessariness of these original goals, there remains a sense in which the behaviours—either automatic, or habitual, or both simultaneously—are nevertheless constrained by the particular shape that they took in response to those early goals. Thus, as Wood and Neal recognize, "...given that habits typically originate in goal pursuit, habit performance often inadvertently promotes goal-consistent outcomes" (2007, p. 847).[142] It is not that this "goal-consistency" picks out an actively endorsed objective to which one's habitual behaviour conforms, but rather, it is merely that one's habitual behaviours bear the mark (so to speak) of some previous goal. This state of affairs leads Wood and Neal to claim that "…habits may be goal directed…even though they are not goal dependent" (2007, p. 847). In other words, habits—insofar as they exemplify automatic behaviours—may not require an occurrent and consciously considered or held goal to give them shape but they may nevertheless betray the structural features of being responses to prior goals; that is to say, they may display a certain telos.

[142] And this conforms to Reason's observation that action slips tend not to betray random or un-patterned behaviours but rather to exemplify recent and structured or commonly engaged in behaviours (1979, p. 161).

There is, however, one important difference to be drawn between habits and automaticity according to Wood and Neal. For them, habits fail to display variability whereas certain *goal-dependent* forms of automaticity appear capable of adaptively adjusting to changes in the environment that are relevant to goal-pursuit. Along with Bargh and Barndollar (1996, p. 461), they see goal-dependant forms of automaticity as operating in terms of a tacit "strategy" for adaptively dealing with a changing environment. Thus, there is a stronger, more active sense in which instances of this particular form of automaticity are shaped by goals (even though this strategic guidance may issue unconsciously).

Another important point about the interface of habits, automaticity, and goals is that both habits and behavioural forms of automaticity tend to develop incrementally over extended periods of time—using one's bicycle to go the grocery store twice in a row does not yet make that behaviour a habitual or behaviourally automatic one.[143] And this gradual developmental process is important, according to Wood and Neal, "…because it insulates habit dispositions against short-term changes in behaviour that occur as people flexibly pursue their goals" (2007, p. 850).[144] What this means is that habits and non-goal-dependent forms of behavioural automaticity tend to be resistant to the occasional departure from the norm. In other words, some unusual environmental feature might once in a while require a change in one's normal or habitual behavioural output in order to reach a given end—but if and as soon as that feature is no longer present, one can expect one's behaviour to return to the standard response. Strictly goal-dependent forms of automaticity, on the other hand, tend to be more flexible and less static than

[143] However, presumably, it would be easier to identify the precise moment at which a behaviour becomes automatic than it would be to determine precisely when we are entitled to consider some behaviour "habitual"—but this detail is of little importance to us. I would imagine that if, after observing an individual responding the same way to a given trigger several times and with only minimal if any behavioural variation in between, most would agree that what is being observed is habitual behaviour.

[144] I would add that this holds for non-goal-dependent forms of automaticity as well.

this because they are mediated by and responsive to the more dynamic features of the environment. In other words, with respect to goal-dependent forms of behavioural automaticity, there exists a sort of reciprocal interplay between a changing environment and automatic behaviours which, despite being automatic, show signs of relevant situational adaptivity. The down-side to the greater stability of habit and non-goal-dependent forms of automaticity is that they may at times conflict with one's active goals, as is often the case with action slips.

Now that we understand some of the ways in which habits and automatic behaviours may both differ as well as overlap, in the next section, we will return to the notion of action slips and why they fail to amount to instances of autonomous behaviour. We will then consider some other important examples of automaticity in various everyday settings in order to get a sense of just how deeply automaticity can undermine one's autonomy

2.6 Automaticity, Dissociation and Autonomy

In section 2.4 we saw that action slips formed a prototypical class of automatic behaviours. Moreover, for many sorts of action slips (most notably those of the 'test failure' variety), it was recognized that what led to the behavioural errors was a failure of conscious control to intercede at what Reason called "critical decision points". Indeed, Reason maintained that to successfully carry out a plan often requires that one be able to consciously recognize and take control of one's behaviour at various important junctions. But the problem, with respect to behavioural automaticity, is that it is characterised by a conscious dissociation from the very (automatic) behaviour at issue (even when that behaviour is goal-dependent). So, if the automatic

behavioural process is operating efficiently (i.e. independently of conscious oversight), there is no good reason to suspect one's conscious attention will be at the beck and call of this process; be it at critical decision points or not. Moreover, as we saw in the Stroop experiment examined earlier, in terms of the percept-judgment category of automaticity, one's conscious attention may remain in proximity to aspects of the automated percept-judgment and verbal behaviour and still be impotent to override the automatic process. In fact, in the case of the Stroop experiment, a restricted range of attention limited to a particular focal region was seen to be required for the automated processing effect to occur. But regardless of that closeness and involvement of conscious attention in the production of the automatic processing delay, participants were nevertheless powerless to overcome that delay.

Returning to the example of the action slip involving an individual who failed to entirely disrobe before stepping into the tub, we might claim, along with Reason, that in order to successfully carry out the plan to take a bath, the individual would have been well served by an awareness of the fact that he or she had not yet entirely disrobed. Such an awareness could have provided the opportunity to finish undressing before entering the tub, which would have been in line with the individual's plan—whereas, getting one's undergarments wet was, let us stipulate, not part of the original plan. If we look at this example through the sort of lens that was provided near the end of the first chapter (in section 1.3), we might suppose that the individual had a raw desire to take a bath and let us assume that, at the outset at least, the individual also possessed the reflective desire and volition that the raw desire be effective. In other words, the agent coherently and—insofar as the terminologically revised traditional view is concerned—autonomously willed to take a bath. Now part of the standard procedure for taking a bath involves entirely disrobing prior to stepping into the tub, this procedure may be considered habitual and it is

certainly the sort of behaviour that most people would engage in with enough frequency and consistency for the behaviour to become automatic. Thus, assuming that the participant's stepping into the tub prior to entirely disrobing was an instance of absent-minded automaticity (as the participant reported), we may begin to consider just how the automated behaviour differs from what we might otherwise treat as autonomous behaviour. As stated, at least at the outset the individual had autonomously (we are assuming) willed to take a bath. However, somewhere between the admittedly short amount of time that it took for the individual to autonomously will to take a bath and the behaviour of stepping into the tub, the individual became somehow dissociated from his or her own bodily activity to the point of failing to recognize that he or she had not yet fully undressed. As a result of this failure of recognition and error on the part of the automated behavioural program, the individual ended up with wet undergarments—which was entirely unintended. Clearly, there are two sorts of problems to be identified here. The first issue is that the automated behavioural program made a mistake but this point is a minor one[145] and it is subordinate to our concern with the next issue. Indeed, of greater importance is that feature of automaticity that is far more universal; namely, the dissociation which it implies. This dissociation appears as an invisible barrier of sorts keeping the conscious and reflective consideration of the individual from playing a role in the production of either some or all of his or her own behaviours for a certain amount of time. The behaviours that an individual may display during these dissociated instances of automaticity might not always conflict with or diverge from the individual's reason or will, but they often do, and the dissociation itself is symptomatic of a sort of incoherence between an individual's deliberate or willed intention and

[145] The utility of such errors lies in their ability to render salient the fact that there are often unintended and undesirable consequences to the dissociation of reflective thought and behaviour that is characteristic of automaticity.

his or her own behaviour that I will argue renders personal autonomy impossible in those moments.

The same issues arise in the example of the individual who went outside to pick up the morning paper to read with breakfast and ended up immediately (and automatically) tossing it into the recycle bin. Again, in this case, it was a dissociation of conscious awareness from bodily behaviour that allowed the unintended error of placing the paper into the recycle bin—without recognizing that the paper had not yet been read—to occur. Of course, rather mundane examples like these—although they demonstrate what is problematic with treating automatic behaviour as potentially expressing autonomy—may arouse little concern for revision. Indeed, one might think that they are simply too insignificant with respect to our usual uses of the notion of autonomy to warrant any major theoretical overhaul since, for instance, they don't appear to be of any great moral concern to the individuals in question nor to any others for that matter. But with very little by way of imagination, I propose, one may discover real-world examples of some of the more pertinent and impactful dangers of automaticity. And it is an encounter with these dangers that in part serve to motivate this project. In the next section, we will consider how the dissociation of conscious oversight from a class of one's behaviours that is endemic to automaticity, as well as the errors that such dissociation may sanction, can lead to rather dramatic and unintended results—results which I suggest urge us not merely to take notice, but which require us to revise our model of autonomy.

2.7 Automaticity as a Problem for Autonomy

Beyond the sorts of 'inconveniences' that automaticity may be responsible for when considered as a rather inconsequential by-product of our daily routines and habits, there exists a number of examples of automatic behaviours that may serve to render much more apparent the real dangers of automaticity as it may impact the health, safety, or wellbeing of one's self and others. In section 1.4 of the previous chapter, I provided the example of an individual who automatically yelled at a mugger to leave his victim alone rather than taking pause to consciously reflect about what might be the most prudent option for response (with respect to both his or her own safety as well as that of the victim). The example was one in which a general personal policy to stand-up for distressed others who were under verbal or physical attack had been adopted at a young age. Over the years, however, this policy eventually became automatic and would result in the immediate issuance of either verbal and/or physical defence of these distressed others. The danger, it was identified, lay in the fact that the automatic and immediate reaction in this type of setting drastically impaired the individual's ability to assess the situation and to determine the safest course of action. Perhaps the mugger might have been paranoid and concealing a gun, and drawing attention to the mugger might have resulted in some kind of overblown retaliation like having him open fire upon the would be defender. Like the earlier example of the individual who intentionally planned to take a bath but ended up getting into the tub prior to completely disrobing, the sort of policy in play here may have been initially adopted from an entirely coherent and autonomous stand-point.[146] Nevertheless, later on down the road, after the behaviours that the policy supports had become automatic, that originally coherent and

[146] Again, I am here only talking about "autonomy" from the viewpoint of the traditional coherence model with the modified and non-hierarchical language. Any mention of autonomy from this point on will refer to the same model until I propose further revisions.

autonomous mindset then became entirely superfluous and was no longer required for the issuance of the said behaviours. So, again, in this example, what we are talking about is a kind of dissociation or incoherence between an individual's conscious oversight of his or her behaviours and those very behaviours themselves. The obvious worry here being that without such conscious oversight and control of one's behaviour, one may 'act' in a way that would fail to conform to one's occurrent will had one been consciously aware of what it was that one was doing. In other words, despite the fact that many instances of automaticity may be historically connected to and in some way shaped by past practice, plans, or policies, the dissociation that such states entail cautions us against assuming that such behaviours are the natural or intended consequences of a presently autonomous mind. But let us examine several other examples in order to paint a clearer picture of just how dangerous this sort of conscious disconnect from one's behaviour can be and just how poor a candidate automaticity is for representing one's autonomous volition.

The domain of sports and athletic training may bear the most obvious connection to the sorts of activities and practices that are known to lead to automatic behaviours; that is, in terms of the repetitive and consistent pairing of situation and specified behavioural response that is typical of athletic practice. Indeed it can take years of effortful practice and deliberately focused corrections to achieve a degree of expertise in many, if not most, athletic disciplines.[147] And along with that frequent and consistent pairing of event type and situational response, as we have seen, the repeatedly practiced movements can eventually become so second nature that one's conscious attention can be focused elsewhere and yet those previously practiced movements may still be executed with flawless precision. In fact, in some instances, it would seem, conscious

[147] See the seminal paper by Ericsson, Krampe & Tesch-Römer (1993) for more on the role of practice in the attainment of expert skill.

attention can get in the way of successful bodily movement in sports—especially when it may slow an athlete's reaction time in circumstances where a faster reaction time than one's opponent is critical.

Despite the potential benefits of acting automatically in certain athletic contexts, there remains a sense in which automated behaviours may also lead to negative and unintended consequences. Take for example an aspiring college basketball player who has spent years honing her skill at performing the layup.[148] Let us imagine that in the middle of a game she acquires the ball and automatically performs the layup that she had practiced countless times previously. Her conscious attention, however, is elsewhere (say, fretting about her poor performance on a recent test that might lower her GPA and get her eliminated from the team). Operating (as she is) automatically for this series of movements she unintentionally strikes an opposing player with her elbow as she jumps to complete the layup causing that player to suffer a broken nose. Now, we may add a number of other details to fill in the picture here; we might say, for instance, that she is the star player of the team and that being ejected from the game for such a move would be disastrous and surely cost her team the game. We might also add that she was neither angry nor vengeful, nor playing with deliberate aggression when the illegal elbow occurred. And because she was not consciously attending to her behaviours as they were unfolding, she both did not recognize the proximity of the opposing player in her way nor could she have stopped the blow from happening.

In this example, our star player might have entered into the game entirely of her own volition—she may have done so with the goal of playing to the best of her ability and she may have even intended, at the start of the game, to perform the layup as she had practiced it should

[148] A basic layup in basketball standardly refers to making a drive with the ball when close to the net and then jumping and releasing the ball in such a way as to make it deflect from the backboard and go through the net earning two points for one's team.

the opportunity arise. However, because she was consciously dissociated from her behaviours during the course of events that led to the injury of the opposing player, I don't believe that we are entitled to claim that she intentionally or deliberately injured the other player.[149] Her awareness, while performing the layup, was disconnected from her bodily movements in such a way as to render those movements divorced from anything like a presently autonomous will. The incident of the illegal elbow wasn't merely an example of a lack of coherence between some active raw desire and a supporting reflective volition; it issued in the complete absence of any occurrent conscious desire—either raw or reflective—to injure the opposing player (or even to perform the layup for that matter). Moreover, knowing that such behaviour would get her ejected from the game and likely cost her team the win, we can safely assume that, had she been mindfully aware of her own movements and surroundings, she would have done whatever would have been necessary to avoid injuring the other player while performing the layup. So not only was her behaviour in this instance not autonomous, but it actually runs counter to what she would have willed and done if only she were consciously attending to what she was doing instead of being consciously preoccupied with something other than her unfolding bodily movements.[150]

Of course, one might attempt to argue that had the illegal elbow not happened, then there wouldn't have been any question about whether or not the layup was part of a series of

[149] Remember, we are stipulating that she performed the layup automatically so there is no question about whether or not she was 'really' dissociated from her bodily movements. And although one might refer to the elbow as accidental it was not the case that she merely saw that she was about to connect with the opposing player and simply failed to react quickly enough to avoid it. Rather, she was utterly oblivious to the movements of her body and her immediate environment in those moments.

[150] It is not the mere fact that an unforeseen and unintended accident occurred as a result of her automatic behaviour that signals to us to treat this case as a non-autonomous one. Rather, the accidental character of this type of example merely serves to highlight the psychological dissociation which impeded the agent's personal control of her behaviour in this instance. It is not the fact that an accident took place, but rather, it is the lack of control entailed by the dissociation that stands in stark contrast to the self-governance implied by autonomy. Accidents alone are not very important to us since they befall those who are acting autonomously as well. They are of relevance to us here only insofar as they are the potential consequences of dissociated, automatic behaviour.

autonomous actions performed by the agent, since we might characterize her behaviour as following from her goal to play the game to the best of her ability. Thus, one might maintain that, in at least some cases, instances of automaticity should be taken to count as component pieces of one's autonomous behaviour. And, given that one endorses the *responsiveness to reasons* or *responsiveness to reasoning* models of autonomy outlined in section 1.1, this perspective might appear sensible.[151] However, in my view, when automatic behaviours end up working out in one's favour, it is not the case that they are then to be treated as examples or expressions of one's autonomous volition. Rather, it is more a matter of happenstance in such cases that one's behaviours turn out to be consistent with one's earlier desires or intentions than it is a case of actively self-governed action.[152] Just as the direction of the wind may at times conform to the sailor's desire that it blow eastward so that he may reach his goal of arriving at shore before sundown, so too might one's automatic behaviours at times turn out to be consistent with one's overarching plans or objectives. But this bit of fortuity does not amount to the personal control by an agent of her behaviours. She may be the author of her movements in the extremely limited sense that it is her body that is implicated in the automatic behaviour that has taken place. But, she may deny having behaved with any intention and resist taking responsibility for the movements of her body while it was operating automatically because those bodily movements were not actively under her control. The problem with treating automatic behaviours as exemplifying one's autonomy—when such behaviours turn out to be in line with one's present objectives—because they permit a consistent narrative outline of what one is

[151] On these views her automatic behaviour may be characterized as falling in line with her reasons for playing the game to the best of her ability or her motives to play to the best of her ability given her other beliefs and desires.

[152] We know that some automatic behaviours are conditioned by goals that were actively held at some prior point in time, so the trajectory of such behaviours may not be entirely a matter of chance. However, that such automatic behaviours conform to one's current objectives would appear to be far more uncertain since one's current objectives might significantly depart from one's previous commitments or goals.

doing, is that it masks the psychological disconnect from one's own behaviour and provides an overly simplistic view of human agency and action. Indeed, from this over-simplified perspective there appears to be little reason not to count breathing or one's heartbeat as autonomously self-governed as well, so long as one maintains the general goal of living.[153] Recall that in section 1.1 I claimed that the notion of self-governance should be taken to mean the acknowledged personal authorship and control, by the agent, of the actions he or she commits. As stated above, our star player in this example may disavow personal authorship of her automatic behaviours apart from that her body was involved in performing them. Also, it must be acknowledged that she was not in control of her bodily movements since she was, in those moments, entirely psychologically dissociated from them. Thus, a more careful examination reveals that her behaviours were not self-governed during the course of the incident even though they might have been conditioned by her at some earlier time.[154]

Next, let us examine a different setting: the industrial workplace. Modern day factories have become highly automated. They make use of individual machines to do in minutes what would have once required numerous human workers hours and sometimes days or more to accomplish. Often, factories will have several of these types of machines form part of a line whereupon, at each step, a number of different and separate aspects of production, assembly, and testing will take place. Although these machines may be highly efficient, there are often certain aspects of the production, assembly, and testing process that require human dexterity and input at various points. Having grown up in an industrial city and worked in a number of factories, I have

[153] Of course, given a certain degree of focused attention, one may come to *regulate* either one's breathing or heart-rate or both simultaneously but, for the most part, we will continue to breathe and our hearts will continue to pump to circulate our blood regardless of whether or not we are at all conscious of these bodily processes; so to consider these processes to be autonomously self-governed is on par with maintaining the absurdity that mitosis or any other physiological process that takes place below the threshold of personal awareness is nevertheless self-governed merely because they are happenings that take place within one's body.

[154] Keep in mind also that those previously conditioned behaviours were merely aimed at the successful execution of an athletic maneuver with absolutely no intention to injure opposing players.

direct experience with these sorts of factory floors. I can also say with confidence and authority that line work (as it is commonly referred to), is an extremely repetitive, monotonous, and boring activity. In some cases, a worker will be required to perform the same series of simple movements more than a thousand times in a single eight hour shift alone. Needless to say, this kind of repetition can rapidly lead to one performing the requisite movements automatically. In fact, in this kind of scenario, allowing automaticity to take over may provide one's only form of escape from the mind numbing tedium of repeating the same simple movements over and over again. It is perhaps no real wonder then that many refer to labouring on the assembly line as "mindless work".

Consider the following example: Tom is a line worker in a factory that produces plastic gas tanks. He has worked in the same position for over a year and his ability to perform the behavioural tasks that befall his position have long since become capable of issuing entirely automatically. His job is to connect the gas tanks that come his way on a conveyor belt to an air injecting hose and several plugs before submerging the tank in water, pressing a button to fill the tank with air, and then looking or listening for bubbles. If there are any bubbles in the water, the tank must be removed from the line and sent to be recycled, and if there are no bubbles, then the tank is removed from the hose and plugs and sent along a conveyor belt to the next station in the line. Now, let us imagine that on a given day, Tom is at work on the line, performing his usual duties for his employer. A gas tank comes down the line, he fits it with the hose and plugs, submerges it, presses the button to fill it with air, and then sends it along to the next station. Let's say that he does this several dozen times throughout the course of his shift. Every now and again, the sound of bubbles signals him to remove the tank and set it aside for recycling. Now let us suppose that Tom has settled into the day's routine enough that he becomes dissociated from his

bodily movements and automaticity takes over for his otherwise conscious oversight and control. Perhaps his mind wanders to the topic of where he will take his next vacation and he begins to imagine different tropical destinations in his mind's eye. He might become so engrossed in this fantasy that his consciousness becomes completely focused upon his visualizations of the sorts of experiences that he might enjoy while in these different locations. All the while, however, he unconsciously (i.e. automatically) sorts the gas tanks with a regular degree of accuracy. As the next gas tank makes its way to Tom, he automatically mounts it to the hose and the plugs, submerges it, and presses the button like before. However, this time, there is a problem with the air that is fed into the gas tank. It turns out that the devise that usually governs the amount of air that is injected into the tank is worn out and fails to restrict the air-flow. As a result, upon pressing the button, the gas tank is filled with more air pressure than it can withstand and it expands before exploding, sending shrapnel flying in all directions, and injuring Tom and several nearby workers.

Like in the previous example, Tom may have come to work entirely of his own volition. He may have decided to put in his usual eight hours instead of taking the day off because he is saving to go on a vacation. In other words, his behaviours may be shaped by a goal and motivated by certain reasons, and these reasons may correspond to certain values that he holds (e.g. the importance of leisure to a contented life).[155] So there is a sense in which his behaviour is in line with one or more of his overarching goals. However, like I stated in the previous example, because he was consciously dissociated from his behaviours during the course of events that lead

[155] One might likewise describe his behaviours as flowing from the coherence between one or more of his raw desires (e.g. to earn a daily wage, to save for a vacation, to be a good employee, *et cetera*) and a reflective desire and volition that such raw desires lead him all the way to performing the tasks that are required by his position. In other words, even on the traditional coherence model, we might initially treat his behaviours as amounting to instances of autonomous action. But, as we will see, each of the traditional approaches to characterizing autonomous agency fails to capture precisely why we should not consider his automatic behaviour to be autonomous; namely, because his psychological dissociation from his behaviours precludes their being self-governed.

to the gas tank's exploding and injuring himself and his co-workers, I don't believe that we are justified in claiming that his behaviour moments before the incident was under his autonomous control. His awareness, while attaching the hose and plugs to the tank as well as while pushing the button was disconnected from his bodily movements in such a manner as to render those movements divorced from any sort of presently self-governing will. Instead, his bodily movements appear to have been the result of a conditioned behavioural script that merely played itself out in the absence of any conscious oversight. Had he been consciously attending to what he was doing, he would have stood a greater chance of noticing that the tank was expanding more than usual and he could have stopped pressing the button to halt any more air from being sent into the tank.[156] The gas tank's exploding and injuring Tom and his coworkers wasn't merely an unintended accidental consequence of his performing the tasks that he was trained to perform. Rather, it resulted from Tom's being dissociated from his own behaviours and environment to such an extent that he not only failed to recognize the problem and act to prevent it from getting any worse, but also that he *could not* have recognized nor reacted to what was unfolding before him while in that state. If one is not even aware of the movements of one's body and what is happening within one's immediate environment, it is hard to imagine in what sense one could be legitimately thought to exert any sort of active control over them.

The final example that we will consider has to do with martial arts and self-defense. For individuals that train in most any of the traditional martial arts, there are a series of basic defensive movements and offensive strikes as well as more elaborate sequences of such moves called *kata* (or "forms"). These basic techniques and forms are practiced with great diligence, repetition and guidance from either high-ranking black-belts or the instructor him or herself

[156] It goes without saying that he would, under normal circumstances, never desire or will to allow the gas tank to explode.

(often called the *sensei*). Commonly, one will not be eligible to advance to a higher belt grade until one has performed several specific maneuvers or forms with expert skill; which often requires repeating these moves in practice hundreds of times beforehand over the course of several months. In fact, I recall the master of the martial arts style that I was trained in once saying that one will not have mastered a movement until one has performed it ten thousand times. As we already know, this kind of repetition can lead to such movements being performed automatically; and, with respect to self-defense, reaching the point where one's movements become automatic is in fact one of the primary goals. Indeed, many of the types of scenarios in which self-defense of this sort is most useful involves sudden and unexpected dangers which can surprise, shock, stun, or bewilder the victim of a threat of aggressive violence. And, when in such states, it becomes much more difficult to think clearly about what is one's most prudent move to make in order to safely escape from harm. Thus, when caught in such situations, it can be an important benefit to be able to rely upon one's conditioned self-defense training to automatically handle the threat—despite one's being caught off guard and being mentally unprepared and unready to engage with an attacker—and enable one to escape.

The example of automaticity as conditioned by martial arts training that I will be considering is drawn from my own experience at a time when I was a young brown belt team sport competitor who travelled and competed in Southern Ontario and the neighboring Tri-State area of the United States. The Karate school that I was a part of had two teams, the "A" team and "B" team (of which I was a part) that would travel around to various karate tournaments to compete against other teams from other schools for the top place in a point fighting system. Each team had several members of different ages and ranks that would compete against teams with

similarly ranked opponents and the team with the most points at the end of a match would move one to the next round in the hopes of reaching first place.

The team of which I was a part generally performed somewhat worse than the "A" ranked team from my school. As the "A" ranked team went up against one of the best teams at a particular tournament, I saw my training partner (another top ranked brown belt) get physically dominated by his opponent from the other team (he was outscored by a significant margin and was left with a bruised ego and a beat-up body). Later on in the tournament, it was the turn for my team to face the team that significantly out-performed our "A" ranked team, and I knew that I would have to face an opponent who just dominated a training partner of mine who would usually out-perform me in training. I was very nervous and afraid, but an amazing thing happened as soon as my bout began: I remember bowing to my opponent prior to our match and then, my field of vision narrowed to a small circle and I could barely tell what was going on in that limited window. I'm not sure if it was a result of the fear, or nervousness, or what exactly, but for the first and only time I experienced what is known as 'tunnel vision'. In this state, my conscious awareness was reduced to this small circle of staggered visual impressions surrounded by complete blackness and all of my other senses were completely blotted out. The few images that appeared to me in this small window were presented similarly to images in a photo-album and I felt no sense of connection to what I was seeing. At the time, I'm not sure the images made much sense to me at all. In any event, the match came to an end, and I came back to a normal state of consciousness and found out that I had somehow won the match by a significant margin. I had to ask the people who watched the match exactly what moves I performed since I had absolutely no memory of any part of the bout other than being stuck in that tunnel vision state. It turned out that one of the spectators had captured the match on a video camera and played the

match back for me on the view screen. Upon watching it on the camera I was amazed not only to see what exactly I had done to score so well (since I had absolutely no recollection of any of my bodily movements while in the bizarre dissociated tunnel vision state), but also the precision with which I performed the moves that I had learned in training. To this day, it remains one of the more remarkable experiences that I have had. I would, however, later discover that the experience is not all that uncommon to people involved in combat sports or real-life combat scenarios and that it is likely the result of high levels of adrenaline production.

One difference between this example and the previous two is that there was no real error or accident involved in this case. I entered into the match with the intention of performing to the best of my ability against an opponent that I was next to certain would out-perform me, and my resulting victory was nothing that I would reject as undesired or unintended by me. However, there is simply no escaping the fact that from my point of view, as the match was unfolding, I was mostly unaware of what was going on (save for perhaps a few unclear images which held no meaning to me at the time), and I had no sense of control over my body nor knowledge about what it was doing as it was doing it. It may sound rather hard to believe or understand for someone who has not undergone a similar experience but, I chose to use this example in part to highlight the fact there exists more than one way in which the dissociation that can occur while one's body operates automatically can come about. My body could have done a number of different things while I was in that state: it could have stood frozen with fear, it could have fled, it could have behaved in a purely defensive manner, *et cetera*. Either way, my conscious will had nothing to say in the matter once I became dissociated. From that state, I felt utterly no control or connection to my body and thus I cannot consider my behaviours while in that state to have been self-governed or autonomous. My behaviours throughout the match were obviously conditioned

by my prior training and conscious goals to reach the best of my ability but, as those behaviours became manifest in that instance, I had no oversight nor could I consciously make any regulatory adjustments to the movements of my body. Therefore, I think that we are forced to conclude that my behaviour was simply not self-governed throughout the course of my being in that particular dissociated state.

Now that we see why it is that automaticity is a problem for autonomy, we may begin to consider how and by what means we might be able to buffer the notion of autonomy from this type of threat. As we came to see in this chapter, one of the most important things to recognize about automaticity is that the dissociation that it entails is partially characterized by either an inability or a failure to attend to one's automatic behaviours. In the relevant literature, attention is often used to contradistinguish automaticity.[157] To put it colloquially, they are often seen as two sides of the same coin. It seems crucial, then, that we develop a better understanding of attention and how it may counteract those instances of dissociated abstraction of which we are now familiar. Thus, in the next chapter, we will examine the progressive development and refinement of some of the research upon, and theorizing about, attention and its role in human mental processing in order to get a better idea of just what function attention might be able to serve in sustaining autonomy.

[157] As we saw in section 2.3 it was used to distinguish controlled from uncontrolled or automatic processes.

Chapter 3

> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatter-brained state…
>
> -William James

3.0 Introduction

In the previous chapter, I presented and considered some of the empirical research on the phenomenon of automaticity. I also compared and contrasted automaticity with related phenomena like automatism and habit. In the final section of the chapter, I provided several everyday—but nevertheless, weighty—examples of the potential dangers of automaticity and suggested that in light of the dissociation of conscious oversight from one's behaviour that is the hallmark of automaticity, we ought not to treat any such automated series of behaviours as expressive of an individual's autonomy. In other words, I advocated the position that instances of automaticity ought to be included in that class of phenomena that are generally taken to undermine a person's autonomy by subverting the sort of self-governance that is essential to genuinely autonomous actions. At the end of section 2.7, I made the claim that what turns out to be lacking in cases of automaticity is, in part, any sort of conscious attention to one's behaviours as they are unfolding.[158] To put it another way, behavioural automaticity is characterised, in large part, by a failure of active attention to what it is precisely that one is doing. And this relationship

[158] And this is consistent with Reason's (1979) recognition (mentioned in section 2.4) that action slips appear to be the result of a failure to consciously attend to one's behaviours—or, to use Reason's terminology: a failure to use a 'closed-loop' mode of control—at what he considered to be 'critical decision points' in the carrying out of an action-plan.

between attention (or rather, a lack thereof) and automaticity is prevalent in the literature. Indeed, attention is often treated as the contrary of automaticity in empirical studies. Thus, studies which take automaticity for their object typically turn out to have something to say about attention as well. This is why attention and automaticity are often treated as two sides of the same coin.

Because attention is commonly opposed to automaticity, and automaticity has been identified as a genuine threat to personal autonomy, it will be important to consider some of the core research and theorizing on attention in order to identify whether or not it may provide important leads with respect to how we may insulate a robust theory of autonomy from the threat of automaticity. Empirical research and theorizing on attention is also important to the project undertaken in this dissertation in the sense that the relationship between raw desires and reflective desires, identified in the first chapter, appears to consist (at least in part) of attention. That is to say, what characterizes the link between these two types of desires is that, when they cohere, their connection seems to involve (and perhaps even requires) a kind of conscious and reflective attention to one's raw desires. And attention research may help to provide a better understanding of that relation as well.

Unfortunately, not all researchers have been entirely clear about precisely what they mean when they engage in attention research. Indeed, in some of the theorizing and experimental studies, there is an unacknowledged tendency by researchers to reduce attention to mere selection. And where this is the case, it can become ambiguous and difficult to determine whether researchers are examining forms of attention that are necessarily conscious, or if they are instead concerning themselves only with selective cognitive processes that may occur below the threshold of conscious awareness. A third issue is that subconscious selection processes can

be responsible for delivering some contents to the attention of conscious awareness while ensuring that other contents are never consciously acknowledged, and researchers may study both of these related elements without being entirely clear or consistent about exactly where the line is drawn with respect to that distinction. Indeed, this sort of confound was recognized by Schneider, Dumais & Shiffrin (1984) when they suggested that, "…attention itself can be automatized [as observed in the] (orienting response)" (p. 20-21).[159] These sorts of ambiguity problems are important to keep in mind when examining the literature. It is especially important for us to be certain of the details here since, in some cases, mere processes of cognitive selection may resemble something closer to automated cognitive operations than they do instances of deliberately conscious attending and recognition.[160]

Another issue within the literature that bears mentioning is that the field of attention research has yet to settle upon a single accepted theoretical model from which to understand the phenomenon of attention. Therefore, despite the claim made by James (1890) in the excerpt at the beginning of this chapter that "every one knows what attention is," an informed reading of the current state of the field reveals that there isn't anything like a broad consensus among researchers about how best to understand or make sense of attention in general terms. For this reason, it will be important to first consider some of the dominant and recent models of attention in what follows—both in order to clarify the ways in which attention has been and continues to be studied, as well as to determine whether or not any particular approach is more well suited than the others to help reinforce a theory of autonomy against the worries generated by automaticity—before transitioning to other important aspects of attention and related notions.

[159] The orienting response relies on automatic processes which serve to direct the conscious attention of an individual in the absence of any deliberate intent.

[160] Recall that, as we saw in section 2.1, there is evidence that automatic human behaviours can also be shaped by automatic and environmentally responsive correction processes. And where automated processes can make corrections and adjustments, they must certainly be able to make selections as well.

There are more theories of attention than space will allow me to adequately consider in this chapter but, in the following section, I will present some of the more influential theories to have shaped the field in recent decades as well as a promising candidate theory that has the potential to reshape the field in the decades to come.

3.1 Theories of Attention

3.1.1 A Preliminary: Divided Attention and Dual-Task Research

Much of the empirical research that underlies the various theoretical positions that will be covered in the following sections draws upon the dual-task experimental paradigm and the measurable interference that such procedures may produce on dedicated attention by virtue of dividing it between different goals. For this reason, it is worthwhile to briefly consider both the notion of divided attention and the characteristic features of the sorts of dual-task experimental research methods typically employed prior to dealing with the various theoretical views in order to better understand how the core data was derived as well as how it helped to shape theorizing.

One of the earliest experimental protocols to reveal that attention may not be as straightforward a phenomenon as it is commonly taken to be was developed by the German physician Hermann von Helmholtz. In general, it would seem that one's visual attention, for example, is tightly connected to the direction of one's ocular gaze. Indeed, Armstrong, Schafer, Chang & Moore (2012) suggest that, "in the visual domain, attention and gaze are typically in register, such that the stimuli lying at the center of gaze, and on the most acute region of the

retina (the fovea), tend to be the focus of attention" (p. 151). However, Helmholtz (1867) was able to show empirically that one's visual attention could be separated from the center of one's ocular focus. He did this by building his own special sort of tachistoscope which consisted of a wooden box that was painted black on the inside with two view holes on one end and a pinhole for light on the other which was located at the center of a stimulus display card. Looking through the view holes, one would focus one's eyes at the pinhole of light located in the center of the visual field. A brief flash of light would then be set off inside the box, illuminating the previously dark stimulus card which surrounded the pinhole for a segment of time too short for the eyes to shift focus—and, if one's eyes moved after the illuminating flash of light passed, it would still not alter the position of the after-image upon one's retinas. The stimulus card was covered in a random assortment of unevenly placed letters and it would later be used to determine perceptual accuracy after the initial experimental exposure. What Helmholtz discovered was that when he focused attention to a specific section of the visual field at a distance from the pinhole of light, while nevertheless keeping his eyes trained upon that central pinhole, he was able to accurately distinguish the stimulus card letters in that section. According to Wright & Ward (2008), "what Helmholtz had demonstrated was that, when their locations do not coincide, visual analysis required for object identification appears to depend on attentional focus more than on ocular focus" (p. 5). This showed that visual attention could be separated from the direction of gaze and it introduced a new layer of complexity that would require going beyond the standard understanding of attentional processes at that time.

According to William James (1890), Helmholtz's discovery was "one of the most important observations for a future theory of attention" (p. 438). However, despite this (as it would turn out prophetic) statement, James maintained that it was not the case that attention was

simply shifted from the center of ocular gaze. Rather, drawing upon the work of Hering, James claimed that attention was shared between the ocular focal point and the intended peripheral region of the stimulus card. Hering's view would later be shown to be false[161] but, it was nevertheless responsible for having led James to conclude that, in addition to one's attention to what is directly in-line with the central focus of one's eyes, objects outside of one's ocular gaze could also be simultaneously "accommodated" by attention. In other words, Helmholtz's work was seen to be the first empirical support for the idea that attention could be concurrently divided between different objects or tracking tasks. Of course, anecdotally, it was widely recognized—long before Helmholtz—that, in some cases, attention may be divided among different objects or tasks at a time; however, perhaps the most significant contribution made here to later work in the psychology of attention was the idea that not only could attention be simultaneously divided between objects or tasks, but that studying attention under dual-task conditions provided a new way of experimentally examining and improving our understanding of the phenomenon of attention.

Although the impact of this new "divided attention" experimental protocol would not be immediately realized, it would nevertheless significantly shape the majority of the research conducted by later theorists.[162] It was not so much the fact that visual attention and ocular gaze, in particular, could be separated that would deeply influence the broader field of attention research, but rather, it was that attention in general could be divided between separate objects and tasks, and that studying attention during these instances could provide new insights that would turn out to be of greatest influence. Indeed, recently, Braun has claimed that, "a number of crucial advances in our understanding of attention are the fruit of divided attention experiments"

[161] See Wright & Ward (2008, p. 6).
[162] See Wright & Ward (2008, p. 12) & Armstrong *et al*. (2012, p. 151).

(1998, p. 345). And it was the innovative and original work of Helmholtz that helped to inspire further research along such lines.

But not all divided attention experiments were restricted—like Helmholtz's ground-breaking work—to solely visual factors. Indeed, shifting away from the emphasis on visual attention, researchers recognized that human beings are, in some cases, able to divide their attention between different tasks or stimuli with respect to other sensory modalities as well. For instance, one of the commonly cited auditory experiences of divided attention is known as the cocktail party phenomenon. This auditory attentional phenomenon is characterized by the experiencer being engaged in listening to one conversation among many others in a crowd containing several separate groups of interlocutors and yet still recognizing when his or her name, for example, is mentioned amongst a different group of people within that crowd—and against a background of otherwise indistinguishable chatter.[163] Perhaps not surprisingly, a number of researchers drew upon this sort of example in constructing different dual-task (i.e. dual-stimulus stream) experiments aimed at studying divided auditory attention. Indeed, the opportunity to present dichotic stimuli in virtue of the distinct physical positioning of the human ears is something that has been widely exploited by researchers.[164]

While it is true that the majority of the experimental literature deals primarily with the visual or auditory modalities, in some cases, individuals have also been shown to divide their attention across these and other sensory modalities. What is of interest to researchers in most every case is where the division of attention leads to interference or a breakdown in the performance of either task. It is this interference between tasks that allows researchers to

[163] The basic fact that a listener could isolate his or her auditory attention to the conversation taking place amongst a particular group of speakers from within a larger crowd of concurrent conversations was itself a problem for attention theorists. See for example: Driver (2001, p. 54) & Baddeley (1990, p.118).

[164] See for example: Broadbent (1954); Treisman (1960).

quantify the effects of divided attention upon behavioural output. Indeed, as Braun notes, "when two tasks fail to interfere, relatively little can be concluded about attention" (1998, p. 339).[165] In contrast, the failure to accomplish one or more of any two simultaneously attempted tasks—ones that can be successfully performed in isolation—is treated as revealing the limits of attention's influence upon behaviour.

There are many different dual-task (or dual-stimulus stream) experimental designs that have been deployed in the analysis of attention. In fact, there are far too many to elaborate upon here. However, in order to get a sense of what is involved in dual-task research, consider the following prototype example adapted from Allport *et al.* (1972): In this experiment, the first task that participants were given was an auditory shadowing task. This task consisted of a one minute segment of prose presented binaurally through earphones that participants were required to repeat word for word as the recording progressed. The second concurrent task to be performed was to sight-read a sheet of musical notes and play that music on a piano. Participants were instructed not to treat either task as primary and not to correct errors but to simply continue as evenly and accurately as possible for the duration of the trial. Both the verbal shadowing speech and the piano performance were separately recorded for later comparison with the actual stimuli to check for performance errors and interference.

In the above example, each task was related to a specific sensory modality and each made performance demands upon the participants that were later evaluated in terms of precision/accuracy. In other experiments, each task might make demands upon the same sensory modality or, rely upon a combination of different modalities. In some cases, the only performance requirement might be a simple follow-up verbal report of attentional recall between

[165] However, see Allport, Antonis & Reynolds (1972) who use two examples of dual-task non-interference to argue against a general purpose limited-capacity view of attention.

two streams of presented stimuli. In other experiments, interference effects might be measured in terms of the reaction times of dual-task participants against their own reaction times from single task performances or against those of a control group. In short, experimental manipulations of the various aspects of dual-task research appear to be limited only by the imaginations of the researchers themselves and the questions about attention that they seek to have answered. Varying the experimental protocol along the lines mentioned above (and in other unmentioned ways) may produce markedly different results. In fact, it is the continuous refinement of the thoughts and questions about attention generated by an ever-changing—and sometimes conflicting[166]—landscape of empirical data that is primarily responsible for the various theories of attention that will be considered below. The following several sub-sections present some of these notable theoretical advances in their chronological order of appearance as new data continued to apply revisionary pressure to earlier models.

3.1.2 Broadbent's Bottleneck Theory

Although attention has been studied as a psychological phenomenon at least as far back as James' (1890)[167] treatment of it in his *The Principles of Psychology*, it wasn't until the late 1950's when Donald Broadbent proposed a "bottleneck"[168] model of attention that research in this area began to garner significant attention. Indeed, according to Moray, "the renaissance of

[166] See, for example: Wolters & Prinsen (1997, p. 764-765).

[167] In fact, the study of attention reaches at least as far back as Descartes if we widen our scope to include philosophical analyses of the concept.

[168] Bottleneck theories of attention are also commonly referred to as *filter theories*. Both terms will be used in the following.

attention theory in the late 1950s was truly radical" (2007, p. 3)[169]; and this is, in part, because psychological research in the West, for the previous several decades, had been dominated by behaviourism[170] and its particular penchant (to put it mildly) for reductive (or even eliminative) approaches to human mental activity and cognition in general.[171] Thankfully, by mid-century, the arrival on the scene of cognitive psychology gave rise to a more theoretically nuanced and experimentally interesting approach to the study of attention, and it was the publication of Broadbent's (1958) *Perception and Communication* that helped to usher in this new way of conceptualizing its study. Broadbent's model, apart from offering a more robust and realistic account of human attention than behaviourists could hope to provide[172], also quickly became the dominant prototype for attention research. And the influence of his thinking upon cognitive psychology can still be felt today.[173] Indeed, there remain a number of current researchers who continue to make use of bottleneck metaphors in theorizing about attention. We would do well then, to consider just what is involved in Broadbent's bottleneck theory.

In order to appreciate the context in which Broadbent's views developed, as well as their implications, we must first acknowledge the importance of the notion of a *limited capacity* system. The notion of a limited capacity system is one that Broadbent appropriated from theorists of his day who were working on engineering challenges faced in the domain of information technology. Indeed, according to Mole, "It was the technology of the telephone

[169] This sentiment is echoed by Kahneman, who claims that "By the end of the 1950s, the situation had altered radically, and the newly legitimized concept of attention was a central topic in an emergent cognitive psychology" (1973, p. 2).

[170] See, for instance: Wright & Ward (2008, p. 11).

[171] See, for instance, chapter 10, section 3 of Dashiell's (1928) *Fundamentals of objective psychology* wherein the author makes the tortured suggestion that attending is reducible to a form of posturing—that 'attention' is equivalent to 'a tension' of the body. Such overly simplistic treatments have been (rightfully) consigned to the dustbin of history.

[172] Indeed, as Kahneman notes "the main function of the term 'attention' in post-behaviouristic psychology is to provide a label for some of the internal mechanisms that determine the significance of stimuli and thereby make it impossible to predict behavior by stimulus considerations alone" (1973, p. 2).

[173] See for example Driver (2001, p. 54 & 56).

exchange that most naturally suggested itself as a metaphor for attention at the time when Broadbent was writing" (2009, section 1.6). At that time, telephone exchanges were the bustling centers of activity where incoming telephone calls would be intercepted or *filtered* (by a telephone company employee) before being rewired to the appropriate channel. Such systems were representative of limited capacity in the sense that there existed a finite number of connections that could be maintained at any one time; and further, once a given channel was occupied, it could not accept any other inputs until the original connection was terminated. Broadbent maintained that information processing within the human brain was subject to similar capacity constraints. For him, the hub of activity for human cognitive processing took place at a single informational bottleneck between two cognitive systems which would normally operate as parts of a sequence. The first of these cognitive systems, the *sensory registration and storage system*, was believed to be capable of handling vast amounts of information from various modalities; whereas, the second system of *perceptual analysis* (the system responsible for delivering information to the attention of conscious awareness and response structures), was believed to have a far narrower informational bandwidth than the first. The central idea behind his model was that a vast amount of information would reach a cluster point wherefrom only certain pieces of data would make it through to further processing by the limited capacity perceptual analysis system and ultimately to conscious attention.[174]

Researchers and theorists who adopted Broadbent's views assumed that it was only once a bit of information made it through this filter phase that it became available to attention and response mechanisms. More precisely, any representational information that successfully made it past the bottleneck was inevitably taken to be information that was attended to and that could

[174] It should here be noted that, in addition to being influenced by ideas in information technology, Broadbent's model was also largely shaped by the results of empirical research that he conducted in a number of dichotic listening experiments. See the entries under Broadbent in references for more detail.

therefore potentially be responded to. There was however, something of a longstanding debate about precisely where to position this supposed bottleneck[175], with some theorists arguing that the bottleneck must be located at the large capacity sensory registration and storage stage (this was Broadbent's view)[176], while others[177] argued that it was more likely to be found at the perceptual analysis stage just before response initiation. Those that maintained that the bottleneck was located at the initial sensory registration and storage stage were known as *early selection theorists*, whereas those that believed the bottleneck to be located at the perceptual analysis stage were considered *late selection theorists*.

Early selection theory was characterized by the following two notions: First, that basic physical properties can be detected without the help of attention (this would be accomplished by the larger capacity sensory registration and storage system); and second, that the more sophisticated semantic properties require attention in order to be registered—this sort of information would be handled by the perceptual system and the attentional capacity at its disposal was taken to be limited to a small number of separate streams of semantic information (if greater than one, typically no greater than two). To illustrate this account of attention, consider the following example: one may automatically notice—without intending to, and simply by virtue of the basic physical property of sound—that one's telephone answering machine is receiving a message in an adjacent room to the one in which one is located. However, in order to determine what exactly is being said by whoever is leaving the message requires that the meaning of the caller's words become the focus of one's conscious attending. Once again, the early selection view maintains that the mere sound of the caller's message is registered by the initial sensory registration and storage system (along with numerous other basic sensory

[175] See for example Hancock, Oron-Gilad, & Szalma (2007, p. 46).
[176] See also: Treisman (1969).
[177] Most notably Deutsch & Deutsch (1963), but see also: Keele (1973) and Norman (1968).

impressions), but the comprehension of any semantic features requires the engagement of the perceptual system which selectively receives its contents only once the larger amount of general sensory data is inhibited at the earlier bottleneck.

Although early selection theory appeared to adequately explain certain observed experimental effects, the model was not free from undesirable consequences. Indeed, one troublesome result of the view was that any unattended semantic properties could not be said to produce psychological effects since they would remain unrepresented and unanalyzed at that level; and this aspect of the view would draw serious criticism, since a vast body of data on priming effects stands as evidence to the contrary of this supposition. The theory would soon also face other experimentally motivated challenges and criticisms.[178] These challenges were initially advanced by late selection theorists who, having conducted further experimental research of their own, found the early selection view to be either lacking in some critical respect or simply ill formulated to explain their findings. It is to this late selection view that we turn next.

Unlike the early selection view of the limited capacity of the perceptual system, the late selection view maintained that the majority (if not all) of the perceptual stimuli that we encounter ends up being automatically processed by a large capacity system.[179] Those who adopted this position argued that the bottleneck's role was not to deny certain bits of perceptual information from being analyzed—but rather, these late selection theorists saw the bottleneck as the gateway between the numerous inputs that would each automatically undergo perceptual analysis and the few of those that would go on to become both conscious and available to memory. Indeed, what characterizes the late selection view is not that certain perceptual stimuli require conscious

[178] See for example: Deutsch & Deutsch (1963, p. 81-2).

[179] Thus, late selection theorists maintained that both the sensory registration and storage system as well as the perceptual system were large capacity systems.

attention in order to be analyzed (as the early selection view maintained), but instead, it is the fact that one becomes conscious of certain perceptual data and that this data is encoded in one's working memory that is the hallmark of attention.

Similarly to the early selection view, however, late selection theory also entails a problematic commitment; namely, that there exists a sharp separation between perceptual items that are analyzed but not attended to, and those that are attended to and remembered. The worry here is that late selection theory maintains that attention is not at all involved in the perceptual analyses that take place prior to the bottleneck; but recent neurological research by O'Connor, Fukui, Pinsk & Kastner (2002) and others[180] reveals that this is not the case. Using fMRI, O'Connor *et al*. (2002) showed that "attention modulated neural activity in the human lateral geniculate nucleus (LGN)" (p. 1203)—an area of the brain that becomes active prior to the cortical processing centers. In their study, attention was shown to impact the LGN by improving neural responses to attended stimuli, reducing neural activity for unattended stimuli, and increasing baseline neural activity in the absence of visual stimuli (p. 1203). What this means is that the influence of attention in fact does appear to reach further down into the pre-bottleneck neural substratum than late selection theory (at least as it was originally conceived), would allow.

As should now be clear, both the early and late selection theories faced some problems in that they each entailed certain commitments that would later be shown to be false. As noted by Mole, a potential source of these sorts of problems for both the early and the late selection views is that they each ostensibly rely upon an "assumption about the *linearity* of the processing stream in which selection occurs" (2009, section 2.1.3). However, if the brain systems underlying attention make use of a parallel distributed processing architecture, then strict talk of "early" *vs*. "late" processes may not only be inaccurate and misleading, but it may also confound theoretical

[180] See also: Armstrong *et al*. (2012, p. 152); Braun (1998, p. 331); & Driver (2001, p. 70).

advancement in the face of conflicting evidence. And, according to Kahneman, studies of divided attention reveal that it is in fact the case that "parallel processing of simultaneous stimuli does occur" (1973, p. 121). Therefore, we can be reasonably confident that remaining bound to the notion of a neat sort of linearity simply won't do. Moreover, the debate between the early and the late selection theorists would ultimately prove to offer little if anything by way of a payoff. Indeed, as Hancock *et al*. (2007) suggest, the debate "never really reached a definitive conclusion" (p. 46). Nevertheless, the notion of a limited capacity that was introduced by Broadbent would go on to influence a number of other theories of attention. In fact, numerous theoreticians recognized how useful it was to treat attention as a limited capacity system and therefore many retained that idea as a central element while constructing their new accounts. At the same time, many of these theoreticians would also attempt to avoid some of the problematic aspects of the filter theory debate mentioned above. Next, we will consider some of these additional accounts, beginning with another widely influential theory proposed by Daniel Kahneman.

3.1.3 Kahneman's Effort Theory

Just as Broadbent treated the notion of a limited capacity system as the cornerstone of his theoretical framework, Kahneman's later effort model of attention likewise regarded the idea of limited capacity to be central to the study and explanation of attention.[181] Indeed, Kahneman even claimed that he intended his model to "compliment rather than supersede models of the

[181] Although he often referred to his theory of attention as a capacity model, I will primarily refer to it as an effort model so as to distinguish it somewhat from Broadbent's earlier account.

structure of information-processing" (1973, p. 11), like the one advanced by Broadbent. The main difference between the two being that, whereas Broadbent's theory attempted to outline the structural architecture of the systems involved in attention, Kahneman's focus was directed more precisely at the "relations of influence and control between components of a system" (1973, p. 11). In other words, Kahneman claimed to be more interested in the mechanisms of selection and their functional roles above their mere structural organization.[182] However, he nevertheless developed a schematic representation of the finer structural layout of the relations between the numerous components and cognitive sub-systems that he took to be involved in attention.[183]

For Kahneman, the fact that the mechanisms of attentional selection existed was without question. Indeed, he noted, for example, the difficulty in predicting whether a pigeon that was trained to prefer a red triangle over a green circle would, on a later trial, prefer a red circle or a green triangle. The question is concerned with whether the animal would make an untrained selection based on the newly separated preferential features of colour or shape. Pigeon behaviour was found not to be uniform on such trials. What this and many other examples led Kahneman to believe was that an organism retains a certain degree of control or choice with respect to what stimuli will influence its behaviour.[184] Indeed, he claimed that, "the organism *selectively attends* to some stimuli, or aspects of stimulation, in preference to others" (1973, p. 3); which renders the contention that attention may be a product of the strictly environmental control of behaviour implausible. Moreover, borrowing a classification from Treisman (1969), Kahneman proposed that there are a number of different sorts of selective activities available to an organism, each of

[182] The three main concerns of Kahneman's effort theory were to develop an understanding of: 1- what is involved in determining task demands; 2- what is responsible for regulating attentional capacity; and 3- how attentional resources are allocated (1973, p. 10).

[183] Although its main components will be described in this section, see his *Attention and Effort* (1973, p. 10) for that graphic.

[184] And this is another reason why an overly-simplistic and purely behaviourist account of attention was recognized to be untenable.

which might potentially operate according to a different set of rules and could possibly be controlled by different mechanisms.[185] These different selective activities were classified in response to what they required of the experimental participant. For instance, the participant might be required to select "inputs (or stimuli) from a particular source; targets of a particular type; a particular attribute of objects; [or] outputs (or responses) in a particular category" (1973, p. 3). And it is important to remain cognizant of these differing task demands since, studies that focus upon different tasks—and that therefore make use of at least slightly different experimental procedures—may produce inconsistent data due to the possibility that they may actually be examining different selection mechanisms.[186]

In addition to his acknowledgement of the varieties of selective attention, Kahneman was also deeply interested in building the case for what he called the *intensive* aspect of attention. Indeed, his careful treatment and development of this aspect of attention reveals it to be a primary and indispensable part of his effort theory. Drawing from common usage, Kahneman claimed that "the term 'attention' also refers to an aspect of *amount* and *intensity* [italics added]" (1973, p. 3). In other words, Kahneman both recognized and highlighted the fact that the amount of attention that an individual may allocate at a given time lies somewhere along a gradient. To illustrate this feature of intensity, Kahneman provided the image of a student who might not fully apply himself to a given lecture. The student may be either tired or, day-dreaming about something unrelated to the lecture. In the first case, because of his drowsiness, he has "less attention to pay" says Kahneman; while in the second, he may be charged with simply attending

[185] And this might help to make sense of some of the conflicting data produced by different experimental studies (i.e. in terms of the supposed location of a possible bottleneck somewhere in the system that was previously mentioned).

[186] This view is consistent with that of the *selection-for-action* attention theorist Odmar Neumann whose views will be considered in section 3.1.5. Worth mentioning here, however, is that Neumann takes this view of multiple selection mechanisms one step further by suggesting that they do not require "the Procrustean bed of a single functional model of attention" (1987, p. 375). In other words, these selection mechanisms might not all be easily explained by an overarching monolithic theoretical construct.

to the wrong things. Another revealing example can be seen in the behaviour of most cooks when they perform complex cooking tasks. For instance, one may easily carry on a conversation with a helper while cooking for a number of guests at a dinner party and yet, one will often temporarily abandon such a conversation when working through the more complicated aspects of the preparation or, when faced with a pot that is boiling over or a dish that is burning or has caught fire. The fact that many will do such things (i.e. briefly put conversations on hold while focusing on more demanding tasks), seems to suggest that people in fact do have a limited amount of attention to pay to a select number of action options and, that some tasks require a greater amount of attention than others. Moreover, in this last example at least, it would seem clear that sometimes, our distributed (yet limited) amounts of attention need to be pooled together and away from less pressing matters in order to successfully navigate immediate and pertinent challenging tasks.

How challenging or complex a task may be, along with its resulting effect upon the intensity of attention elicited or required, is something that Kahneman credits Berlyne (1960) for first noticing. In addition to complexity, Berlyne identified "novelty" and "incongruity" as further elements responsible for rendering certain stimuli more attention grabbing than others. However, Kahneman recognized that Berlyne's research was centered primarily upon involuntary aspects of attention. As a result, Berlyne was mostly occupied with the "level of arousal" that various stimuli might elicit rather than the degree to which one voluntarily allocated one's attention.

Because Kahneman was interested in developing an account of voluntary selective attention, he had a different idea of what might impact the intensity of attention; namely, he maintained that, "in voluntary attention the subject attends to stimuli because they are *relevant to*

a task that he has chosen to perform [italics added], not because of their arousing quality" (1973, p. 4).[187] Clearly then, Kahneman was more concerned with the individual's own cognitive contribution to the degree of attention maintained than he was with the mere influence of environmental factors.[188] The primary reason behind distinguishing between the involuntary and voluntary aspects of attentional intensity and its regulation, for Kahneman, was to disentangle his understanding of the effortful deployment of attention from the broader and less specific notion of arousal. Indeed, he claimed that "the intensive aspect of attention corresponds to *effort* rather than to [the] mere *wakefulness* [italics added]" (1973, p. 4) that is commonly denoted by arousal. Put differently, Kahneman maintained that "the effort that a subject invests at any one time corresponds to what he is doing, rather than to what is happening to him" (1973, p. 4).[189] Importantly then, Kahneman was not just concerned with the intensity of attention, but he also considered it often to be something actively engaged in by the individual; whereas he seemed to treat the intruding influence of certain external stimuli as (at least initially) a passive affair—i.e. as something that foists itself upon the individual without warning or relevance to current plans and intent.[190]

---

[187] As will become clear in section 3.3, this sort of reference to personal choice is rarely seen in attention research.

[188] However, that is not to say that Kahneman chose to ignore the involuntary effects upon attention of environmental factors. Rather, his aim was to integrate the understanding of the intensive aspects of voluntary attention with those involuntary effects previously studied.

[189] Notice how this parallels (and predates by nearly two decades) the quote by Velleman in section 1.1. Problematically, Kahneman also makes the apparently contradictory claim that, "in general, we merely decide what aims we wish to achieve; the activities in which we then engage determine the effort we exert" (1973, p. 14). But the tension between these two statements may be resolved if we take the word "determine" to mean "set the level of" or "constrain the amount of". However, Kahneman also claims that "the mobilization of effort in a task is controlled by the demands of the task, rather than by the performer's intentions" (1973, p. 17). But if the "performer's" intentions play no role in the "mobilization of effort", it remains unclear in what sense the effort invested amounts to something *he is doing*, instead of something that is *happening to him*. Kahneman did not seem to notice this tension between his claims.

[190] However, he does suggest that the influence of such factors may be related to an individual's "enduring dispositions" (1973, p. 4). Notice the parallel to Reason's suggestion that *previously conditioned motor programs* may take over during actions slips mentioned in section 2.4. In both cases, when one's active attention is not governing one's arousal or behaviour, one's level of alertness as well as what one does seems to fall back on previous (perhaps more entrenched) selections for action and response.

Returning to the central idea shared by both Broadbent and Kahneman that an individual's attention is constrained by a limited capacity system, we find that whereas Broadbent maintained that there was a rather specific structural feature that was responsible for the limitation of attention (namely, because different sensory items were competing to bias a single mechanism in order to gain access through a single channel), Kahneman's model, on the other hand, advanced a less restrictive view. Indeed, for him, the competition between potential items of attention was characterised not by appeal to a single monolithic mechanism, but rather, by the demands on attention that either of a number of potential tasks required.[191] In this way, the effort model of attention was able to make some sense of how an individual's attention could sometimes function adequately even while divided between tasks.[192] Another feature of attention that appears to be better accounted for on the effort model has to do with the "variations in the difficulty of what a subject is trying to do [which] are faithfully reflected in variations of his arousal level" (1973, p. 9). Recall the cooking example provided earlier wherein a dinner party chef temporarily withdrew attention from a discussion while facing a particularly demanding aspect of the cooking task or while handling a more pressing concern such as a pot that is boiling over or a dish that has caught fire. Such behaviour reveals that there is more to attention than the mere filtration of select captured data and its eruption to the level of consciousness.[193] Instead, this and similar behaviours suggest an extra feature of intensity or effort that is limited like

[191] It goes without saying that the pertinence of a given task to the individual's objectives also plays a mediating role but this is tacitly taken to be true of the bottleneck view as well.

[192] See for example, Allport *et al*. (1972).

[193] Indeed, drawing upon empirical research, Kahneman argues that, "attention operates by *emphasis* rather than by *filtering* [italics added]" (1973, p. 134).

attention is[194] and that functions in collaboration with the simpler selective aspects of the cognitive processing of stimuli.

For Kahneman, this extra feature of effort fits into his theory of the attentional system in the following way: first, at any given time, there are a number of possible activities that are open to an individual; these activities are each evaluated in light of the demands that they make upon available attentional capacity (i.e. in light of how much effort they require). The available capacity itself is something that is influenced by various determinants of the individual's arousal level (greater arousal tends to increase capacity enabling the individual to invest greater attentional effort; while low levels of arousal, on the other hand, reduce available capacity and diminish attentional effort). The final main component—the "allocation policy", is influenced by the individual's intentions or enduring dispositions, and the capacity demands of the tasks that may be performed. Together these operate to determine which of the possible activities to pursue and how much effort to invest.[195]

The above paragraph describes, in outline, Kahneman's effort model of attention which was both shaped in response to the challenging findings of empirical research on attention performed in the wake of Broadbent's filter theory and developed to remedy some of the troubles of that earlier view. But Kahneman's model was not the only proposal to address those findings and problems. Indeed, another theory known as the multiple-resource model would also turn out to be a major contender in post filter theory attention research. It is to this account that we turn to next.

---

[194] In fact, Kahneman says of this feature (or, as he calls it: "a nonspecific input"), that it "may be variously labeled 'effort,' 'capacity,' or 'attention'" (1973, p. 9), effectively blurring any real distinction between his use of these terms.

[195] Kahneman also notes that "there appears to be a rule that when two activities demand more capacity than is available, one is completed" (1973, p. 11), which suggests that the allocation policy is biased towards successful completion of one task at the cost of failing to complete another competing task, rather than evenly distributing insufficient effort to complete either of the two tasks.

3.1.4 Wickens' Multiple Resource Model

As was made clear in the previous section, Kahneman's widely influential effort theory of attention was both inspired by and borrowed some key notions from Broadbent's earlier model. Most importantly, it retained the idea of attention as a limited capacity system. Similarly, Wickens' also widely influential multiple-resource model (MRM) of attention was inspired to a significant extent by the work of Kahneman—especially in terms of his treatment of attention as a sort of energetic resource (i.e. in terms of the intensity component). Indeed, Wickens claims, "I have always been heavily influenced by a dominant theme of Kahneman's (1973) book: the association of attentional resources with a mental effort that can be allocated to tasks" (2007, p. 243). The focus upon both attentional resources and multiple task demands is something that would greatly shape Wickens' theoretical approach. His interest in and emphasis upon these two components, led his research to extend beyond the laboratory and into applied aspects of the study of attention and its impact on things like driving and piloting behaviour. As a result, his multiple-resource theory of attention has made a significant impact in the field of human factors and upon the engineering of interface designs for complex systems that require human operators. One of the obvious strengths of the MRM, then, is its applicability to practical concerns; but this doesn't yet tell us what makes it superior to Kahneman's earlier view—or if in fact it is, in any way, superior to it. Addressing that question requires that we look at the differences between the two views as well as Wickens' reasons for rejecting certain aspects of Kahneman's earlier model—or instead, his reasons for going beyond the initial account advanced by Kahneman.

One of the main points of divergence for Wickens' multiple-resource view from Kahneman's original effort theory is that, whereas Kahneman seems to have taken a more

parsimonious approach by conceiving the attentional resource to be a single undifferentiated repository, Wickens, on the other hand, discovered reasons to posit several different and often unconnected reservoirs of attentional capacity. And just as the late selection theory was formulated in light of newer research that wasn't entirely consistent with early selection theory, so too is Wickens' argument for a multiple-resource view motivated by empirical data that doesn't appear to sit well with the effort theory's treatment of attention as being accounted for by a single undifferentiated reservoir.[196] The relevant experimental findings come from dual-task studies which led Wickens (1984) to identify four different types of phenomena that, as mentioned, don't fit comfortably with a single-resource account.

The first of these phenomena is what Wickens calls "difficulty insensitivity". On the single-resource view, one would expect that as a primary task becomes more difficult (i.e. as it ostensibly requires more effort or attentional resources), one's performance on a secondary task would begin to suffer as a consequence. And this seems reasonable since, drawing a greater amount of resources for one task from a singular pool shared by both tasks ought to result in the secondary task having less attention to make use of (at least, given that a certain consumptive threshold is reached by the primary task). However, in a number of studies, it turns out that varying the degree of difficulty of a primary task does not lead to any difference in the amount of disturbance or errors produced on a secondary task.[197] This suggests that in at least some cases, the different tasks that are engaged in concurrently during certain dual-task studies might be drawing upon different and independent attentional reservoirs.

[196] It is important here to note that there is no meaningful distinction between Kahneman's use of the term "effort" and Wickens' use of the term "resource". Wickens merely preferred to frame attentional capacity in terms of resources rather than effort since he believed that, "…*effort* suggests a motivational variable that may (but does not necessarily have to) correlate with the commodity enabling performance" (1984, p. 67). Likewise then, in this dissertation, the terms "effort" and "resource" may be treated as referring to the same thing.

[197] See Wickens (1984, p. 76) for a list of studies that support this contention.

Another bit of evidence that is suggestive of the multiple-resource view has to do with the phenomenon of "perfect time-sharing" sometimes observed between different concurrent tasks.[198] Perfect time-sharing is revealed by instances of dual-task performance wherein a participant simultaneously performs both tasks equally as well as each can be done on its own. Here Wickens (1984, p. 76) draws upon several different studies including the previously mentioned example from Allport *et al*. (1972) wherein participants were shown to be capable of sight-reading music and performing an auditory shadowing task at the same time and with the same degree of competence that they could display while performing each of those tasks one at a time. On a single-resource model, engaging in both of these tasks at the same time should result in poorer performance on at least one of them since they would each be drawing from the same attentional reservoir (so to speak). But the MRM can easily accommodate such findings, since it maintains that there are a number of non-overlapping sources of attentional capacity that an individual can make use of synchronically.

The next kind of evidence in support of the multiple-resource view comes from what Wickens labels "structural alteration effects". These sorts of effects are observed in studies wherein the difficulty of one of a pair of unrelated concurrently performed tasks is held constant but is nevertheless changed in some way; for instance, in terms of input modality (e.g. from visual to auditory), or response modality (e.g. from pressing a button to providing a verbal response). In such studies, the single-resource view would predict that so long as the difficulty level is held constant, then there should be no change in the degree of interference registered upon the secondary task. However, a number of studies show that such changes can in fact have an impact upon the degree of interference undergone on the secondary task. Moreover, the MRM of attention can explain such results since the initial task pairings may rely upon independent

[198] Again, see Wickens (1984, p. 76) for references to the supporting empirical research.

resource pools but changing the input or output modality may result in both tasks then relying upon the same resource, which would certainly be a reason for the noted interference effects to register a change even though the difficulty level of the primary task was unchanging.

The last of the phenomena that Wickens mentions in support the MRM is what he refers to as the "uncoupling of difficulty and structure". This phenomenon is elucidated by dividing an initial set of tasks (one of which is rated as being more difficult than the other) and separately pairing each of them with a third task. In some instances, the initial task that was rated as being more difficult actually produced less of a disturbance upon the performance of the third task than did the easier of the original two tasks. Of course, this is not what would be expected on a single-resource view. Rather, on the reasonable assumption that the initial more difficult task simply requires more effort or attentional resources, the single-resource view would maintain that it should cause greater interference when paired with a third task than the easier task would produce—but, according to Wickens, this not always the case. Here again, the MRM of attention appears capable of readily explaining such an occurrence. Indeed, on the multiple-resource view, the differential—and counter-intuitive on the single-resource view—disturbance to the third task can be explained in terms of shared resources *versus* independent resources. That is to say, according to the MRM, the fact that the more difficult task is seen to produce less of a disturbance upon the third task than the easier task does can be explained by the fact that the more difficult task may simply be making use of a different resource, whereas the easier task may draw upon the same resource as the third, which would result in the observed greater interference.

In response to the worries generated for the single-resource view by the sorts of dual-task findings mentioned above, Wickens developed a three-dimensional "box model" measure of

attentional resources and their effects in order to more accurately map the various components and interactions of the human attention system. The first parameter of Wickens' model is concerned with the different "stages" of processing. It includes the encoding, central processing, and responding stages. According to the elaboration of Wickens' MRM provided by Hancock *et al*., this parameter is "essentially a temporal axis reflecting the concern for the sequence of encoding, decision making, and response, which is made in seriatim in traditional stage models" (2007, p. 47)[199], and which represents the different points at which task interference may occur. The second parameter has to do with what Wickens calls the "processing code" which is divided between the spatial and verbal aspects of a task. Next there is a parameter that records the "processing modality" which typically includes both auditory and visual sensory channels.[200] Finally, the types of responses that an individual may generate were separated into either manual or vocal behaviours (the former being associated with the "spatial code" and the latter with the "verbal code").

The above paragraph provides a rendering, in outline, of the heuristic model deployed by Wickens to make sense of the dual-task data and to predict the standard effects of various task combinations upon human attention. In short, it represents one of his major contributions to a multiple-resource view of attention. Worth noting, however, is that Wickens himself admonishes us not to invest too much into paradigmatic models of attention. He cautions: "do not become paradigm bound," and he advocates instead that researchers, "examine the manifestations of the examined phenomena in real-world behaviour, with all of its multitask complexity" (2007, p. 247). In that spirit, the final model that we will examine (and that we turn to next), which also represents a more recent way of conceptualizing human attention, appears to place one of the

[199] They also note, however, that Wickens did consider the possibility of parallel processing.

[200] Although, in principle, one might include any of the five senses available to humans, as well as proprioception.

central "real-world" aspects of attention at the forefront; namely, it recognizes that in normal contexts, attention is intimately related to *action*.

3.1.5 Attention as Selection for Action

At the end of the previous sub-section, we saw that Wickens encouraged researchers not to become "paradigm bound" and to instead turn to examining the phenomenon of attention as it occurs outside of the lab; regardless of how intricate such examples may turn out to be. The thought appears to be that ecological validity and applied research ought to trump theory construction and research projects aimed only at finding support for a given model. And one might agree with the pragmatic counsel that this impetus provides, at least where theory struggles to accommodate various seemingly inconsistent bits of data (such as would appear to have defined the history of the field of attention research).

Turning to real-world examples, then, it is immediately noticed that one of the most salient features of the more commonplace types of attention to occur outside of the laboratory is that such instances are typically concerned with one or more *actions*. A number of researchers have recently chosen to reconsider the role of attention in light of this recognition and their changed vantage point represents a considerable break from how attention research has been conducted and conceptualized previously.[201] Although a number of these researchers maintain that the *selection-for-action* view of attention is in the fledgling stage (i.e. that it is far from providing a thorough account), the view has nevertheless created something of stir, since it does not take for granted one of the fundamental assumptions of the earlier proposals—namely, that attention may

[201] See, for example: Allport (1987); Neumann (1987); and Hommel (2010).

be explained primarily in reference to a limited capacity system.[202] Instead, along with Allport, proponents of this more recent approach recognize that, "the concept of a central limited-capacity system has exercised [a] hypnotic hold on theorists of 'attention'" (1987, p. 410)[203], and they are beginning to challenge this long-standing way of thinking about things.

It is perhaps no surprise, then, that one of the common features of each of the various accounts of attention outlined in the previous three sub-sections was the centrality of the notion of a limited capacity system, be it differentiated—as the multiple-resource model maintains—or not. As Hommel (2010) puts it: "Most of the grand, influential attentional theories have considered attention as a mechanism that administers and organizes scarcity" (p. 122). In other words, attention is standardly conceived as something that the individual possesses in short supply. Interestingly, however, Neumann (1987) notes, "the lack of any physiologically established limit on the information that can be picked up [by the brain] at once" (p. 362). Perhaps, then, as the selection-for-action proponents maintain, the apparent limitation of attention is not in fact a property of the structural features of the brain but, rather, a consequence of selection processes and the action-plans adopted.[204] Such a proposal is certainly consistent with the empirical findings. Indeed, in support of this contention, Allport (1987) reminds us that, "in practice, the *observable* criterion for successful 'attention' to (or awareness of) an environmental event invariably turns on the ability of the subject to *act* voluntarily, or arbitrarily, in response to that event either at once, or subsequently in 'recall'" (p. 408).[205] So when it comes to experimental measures of attention, it seems that, whether or not we are thinking in terms of a limited-capacity system, we are always also taking account of some action(s). Furthermore, once

[202] See: Allport (1987, p. 411).

[203] See also: Driver (2001, p. 56), who suggests that Broadbent's views have been "too influential" and have stifled other ways of thinking about attention.

[204] This is precisely what is persuasively argued in Neumann (1987).

[205] Voluntary acts are intentional or conscious whereas arbitrary acts may be treated as automatic.

the presumption of explaining attention by way of a limited capacity system is abandoned—or at least no longer treated as its only central and defining characteristic—and focus is instead placed upon how attention facilitates everyday action (i.e. what it enables an individual to do), researchers appear to be less bound by the earlier theoretical constructs and closer to a functional understanding of the role of attention as it manifests in the daily activity of regular people.

According to selection-for-action proponents, one of the more serious worries for the capacity limitation models of attention—and a motivating reason to begin to consider alternative views—is, "…the failure of th[e] concept to provide explanatory power" (Allport, 1987, p. 411).[206] One issue is that, in a sense, invoking the theoretical device of a limited capacity system in order to explain the apparent fact that human attentional capacity is limited seems to amount to nothing more than a tautology. Indeed, as Neumann argues, "at the conceptual level, the problem is that this capacity concept can easily be used to produce pseudo-explanations that are in fact mere translations of findings on attention into the language of capacity" (1987, p. 364); but to simply re-describe the observed facts about attention is idle. Another worrisome issue that is contributing to the difficulties for capacity limitation views raised here is that, in most empirical studies, "the notion of 'attention' is generally left undefined" (Allport, 1987, p. 408)[207]; and this lack of clarity may be partly responsible for researchers' failure to recognize when they are merely restating their observations in different terms rather than explaining things.

One of the ways to overcome the explanatory shortcomings of capacity views, according to selection-for-action proponents, is to focus instead upon another of the basic features of attention; namely, its selectivity—and to try to explain what is involved in such discriminative

[206] See also Franconeri, Alvarez & Cavanagh (2013) who claim that, "unfortunately, words such as 'capacity', 'resources', and 'load' relabel the effect without explaining why it occurs" (p. 134).

[207] Hancock *et al*. (2007) echo that this is a problem for theory as well when they suggest, for instance, that, "the multiple-resource model is, strictly speaking, only an architecture. It does not tell us what attention actually is" (p. 48).

processes and how they operate (i.e. to provide a functional account of the selectivity of attention). However, an awareness of the selective aspect of attention is not entirely new to attention researchers.[208] Indeed, as Neumann asserts, "the selectivity of attention has traditionally been viewed as its second major attribute, besides limited capacity" (1987, p. 373). Moreover, he insists that it wasn't that capacity theorists ignored the selective nature of attention, but instead that, "they regarded it as a secondary consequence of limited capacity" (1987, p 373), and thereby, failed both to elaborate much upon the role of selection in attention and action control[209] and to recognize just how important the feature of selectivity is to a thorough account of attention.

In response to the arguably undeserved favour accorded to the notion of limited-capacity, selection-for-action proponents like Neumann propose to reverse the roles of the primary elements of the traditional view and instead treat, "limited capacity as a necessary by-product of the solution of selection problems" (1987, p. 374). By placing the selectivity of attention at the forefront and treating limited capacity as a consequence of selective operations, many of the earlier ways of looking at attention are transformed. For one thing, the earlier view that unselected stimuli are subject to no further processing once a target stimulus has been selected—since this would be too costly to the limited capacity system—is seen to no longer necessarily follow. As Allport emphasizes, "the really important point to recognize…is that selection, in the sense of selective *cueing*, in no sense logically entails [the] rejection or exclusion of the noncued information from further *processing*" (1987, p 409). In other words, higher level operations upon non-selected stimuli are no longer seen as being constrained by a limited capacity. Instead,

[208] As was seen in section 3.1.3, Kahneman considered the selectivity of attention in some detail but his model was nevertheless centered upon the notion of effort/capacity.

[209] Most capacity theorists remained content to treat the selective aspect of attention as the simple allocation of resources (see: Neumann, 1987, p. 374).

attention is constrained by the selective arbitration between, and the carrying-out of, action-plans—but this tells us nothing about whether or not or to what extent unselected stimuli receive further treatment. And, if this is correct, then the longstanding debate between the early and the late selection theorists of the filter theory view may have been the result of a mere erroneous over-emphasis upon the concept of capacity.

From the perspective that selection-for-action plays a fundamental and paramount role in limiting attention there also emerge new issues to be addressed. According to Neumann, the two primary concerns for this view are, "the problem of *effector recruitment* (Which skills, related action goals, are given access to the effector system?), and the problem of *parameter specification* (Which of the possible specifications of an action's parameters is put into effect?) [italics added]" (1987, p. 386). With respect to the problem of effector recruitment, Neumann argues that a certain degree of motivation along with a conducive environment are simply not enough to specify which effector systems[210] are activated by which skills and action-plans. And this is because, in some cases, one's motivational level and the objects in one's environment might lend themselves to a number of different action-plans, each of which themselves may be accomplished by a number of different skillful means. One obvious constraint upon selection here is that one cannot simultaneously perform behaviours that are mutually exclusive since these would require the concurrent activation of the same effector mechanisms (e.g. one cannot clap one's hands while swinging a baseball bat at the same time, since only one of these behaviours can be carried out at any given moment). Another obvious potential constraint would be the action-plans *adopted by the individual* but Neumann skirts this issue since he believes it

[210] Effector systems are cognitive mechanisms responsible for initiating different types of behaviour. They are also basic components of the selection-for-action view of attention.

would introduce the difficult to quantify idea of intentionality[211] (see Neumann (1987), p. 375). Finally, Neumann argues that the apparent capacity limitations of attention may simply be the result of "ongoing action [which] inhibits all other possible actions" (1987, p. 378). Thus, for an individual to be engaged in any occurrent action may itself act to constrain that individual's behavioural activity such that it remains consistent and thereby represses any inconsistent behaviour. In other words, once an effector is recruited to a particular action-plan and engaged in carrying it out, it may then be generally unavailable to other attentional or behavioural operations. One might say that, once a selection is made it seems to stick.[212]

The problem of parameter specification, on the other hand, presents a different set of challenges. First, one's surroundings may not provide the kind of information required to carry out an action-plan (e.g. imagine trying to break open a moving piñata while blindfolded and after having been spun in a circle several times). One might also lack a certain ability or sub-skill to accomplish some action (e.g. one may lack the balance needed to walk across a narrow log bridge without falling into the creek below). Both of these examples amount to instances wherein the action is *underspecified* by the environment or one's skillset. It can also happen that one's actions are *overspecified* by the information present in the environment (e.g. one may plan to have some ice-cream but the thirty-two flavours on offer at the local ice-cream parlour are too many for one to eat at once). According to Neumann (1987), both of these mentioned aspects of parameter specification require selection mechanisms in order to guide behaviour effectively (p. 376). Moreover, he argues that, "parameter specification demands that each parameter is given exactly one value at time [since] a movement cannot go into different directions at the same time" (1987, p. 383); and this is similar to the mutual exclusivity constraint on action mentioned

[211] More will be said about this issue later in section 3.3.

[212] To be fair, however, Neumann (1987) does make note of the flexibility of conscious actions and therefore leaves room for certain pertinent inputs to make intrusions and modify one's behaviour and plans.

with respect to effector recruitment. In both cases, Neumann argues that mutually cancelling behaviours, whether they are the result of overlapping skillsets or effector activations, simply cannot obtain concurrently. The implication for instances of dual-task performance, according to Neumann, is that, "concurrent actions should therefore be possible only to the degree that they can use separate skills [or effectors]" (1987, p. 383). And he maintains that this view provides a better explanation of the dual-task data than has been advanced by the other capacity models of attention since it does not require the added component of resources which he thinks fail to capture the specific nature of interference effects.[213]

The problem of overspecification, according to Neumann, can be split into two sub-problems; namely, the "consistency of selection" and the "continuity of action" (1987, p. 384). He provides the example of picking an apple hanging among several others in a tree. At first glance, this may not appear to pose much of a problem. Indeed, it seems that all one needs to do is to identify the most appealing apple and then to grab it. However, Neumann alleges that things aren't quite that straightforward. For instance, different properties of the apple (e.g. size, colour, distance from one's body, *et cetera*) might serve to specify different selection-for-action parameters, and each of these may be distributed across different apples such that they may each specify different actions and yet, one cannot grasp at more than one of the available apples at a time. For example, one apple might have the optimal size, while another may possess the optimal colour, and a third one still may be at the optimal distance for an easy retrieval. This potential distribution of desirable properties reflects the problem of the *consistency* of the selection—that is to say that, once a selection is made, the choice must remain the same (i.e. consistent) for the action to be completed successfully. A related problem is that, given the distribution of desirable features among the apples, the one that appears to be the most desirable may fluctuate from

[213] See: Neumann (1987, p. 365-6).

moment to moment. For instance, after making a selection, let us say a gust of wind sways the branch revealing a previously hidden apple with a more appealing colour or moves the branch closer to render picking another apple easier. With respect to this sort of case, what we have, according to Neumann, is a problem concerning the *continuity* of action. That is to say that, once an action is initiated, one must stay the course of that action in order for it to accomplish one's goal. Although there are surely some more complicated cases than the apple picking example provided above, Neumann opts for a rather simple solution to at once address both of the identified problems[214] by maintaining that, "in order to solve the overspecification problem, an animal selects one of the competing objects by directing itself towards its position in space" (1987, p. 385)[215]; so it is not just selection-for-action but also an action itself that helps to constrain selection and mitigate the kinds of problems mentioned above. In fact, it would appear to be the case that there is a kind of reciprocal influence at work here: the action of directing oneself toward some object helps to constrain one's selection to that object alone and, one's selection of that object in kind helps to constrain one's further action with respect to it.[216]

Now that we have considered several of the more influential theories of attention to have been advanced over the latter half of the last century, it is important to consider some of the shared issues arising in part from the standard empirical protocols that have informed each of these theories to date. The first of these issues to be examined in the following section is the connection between attention and memory.

[214] Neumann (1987) offers solutions to each of the problems that arise in light of a selection-for-action view outlined in this section but an exhaustive treatment of them is beyond the scope of the dissertation.

[215] This might sound disturbingly close to Dashiell's behaviourist proposal involving "posturing" mentioned in section 3.1.2, but Neumann is only suggesting that physically orientating one's body towards an object may help to solve the overspecification problem; he is not suggesting that attention, in general, is reducible to mere posturing or bodily orientation.

[216] It may be argued whether or not mere orientation towards some object is sufficient to specify an action parameter in all cases but my objective here is not to challenge Neumann's proposed solution; rather, I merely intend to provide what he considers to be one potential solution to the problems identified in this section.

3.2 Attention and Memory

When examining the research and theorizing on memory, a few things that become immediately obvious are that: 1-there are numerous ways of thinking about memory; 2-there are also many ways of operationally defining it; and 3-one may characterize its essential features in different ways. Indeed, in the literature, the term 'memory' has been preceded by a plethora of adjectives, including: episodic, iconic, short-term, working, procedural, implicit, explicit, semantic, personal, direct, long-term, *et cetera*. For our purposes, we need not review such an extensive list. Instead, we can treat memory as being essentially characterised by a single distinction between its short-term and long-term aspects.[217] Short-term memory (where rehearsal and maintenance is absent) is typically believed to be constrained to dealing with a limited number of items (normally less than ten), and to only be sustainable for a brief period of time (often on the order of mere seconds). Long-term memory, on the other hand, appears capable of sustaining an almost unlimited number of items nearly indefinitely. In addition to this basic distinction between short-term and long-term forms of memory, it should also be noted that we will be primarily concerned with *explicit* forms of memory (i.e. memories that can be accessed either verbally or in some other manner by an experimental participant) rather than with *implicit* memory (i.e. experimentally observable signs of memory influences of which the participant remains ignorant) since the bulk of the relevant attention research deals with explicit responses to stimuli.

The vast majority of the dual-task experimental work to have shaped the theorizing on attention mentioned in the previous several sub-sections has also inadvertently tapped—and

[217] Many of the other forms of memory just listed are also commonly treated by researchers as falling under this more general classification.

potentially been conflated with—short-term memory. One of the worries that this generates for models of attention is that, for instance, the apparent 'capacity limitations' that have been attributed to attentional mechanisms may in some cases be the result of memory related factors instead of the supposed attentional constraints. While it is clear that theorists were aware of the connection between attention and memory—recall that the late selection theorists mentioned in section 3.1.2 maintained that attention was characterized both by the fact that one becomes aware of some stimuli and that it is then *available to memory*—whether or not sufficient pains were taken to ensure that attention and memory were kept from being conflated under conditions of empirical study remains ambiguous. Moreover, this ambiguity tends to undermine confidence in the conclusions about attention drawn from such studies.

There is now hardly a doubt that "attention and VWM [visual working-memory; i.e. short-term memory] are *intertwined cognitive operations* [italics added]" (Fougnie & Marois, 2006, p. 533). Indeed, just as attention to an item may influence an individual's ability to subsequently remember it, according to Chanon & Hopfinger (2008), "item memory [also] affects the allocation of attention, influencing both the guidance of attention and subsequent dwell time" (p. 325). So the influence between attention and memory, it seems, operates in both directions. Additionally, in some cases, attention and short term memory can be shown to interfere with one another.[218] But, perhaps most problematic for un-careful attention theorists is the recent evidence that short-term memory may be subject to capacity limitations independently of any influence from attentional processes.[219] All of these recent findings present a challenge to theorists who would gloss the distinction between attention and memory in their experimental designs. The problem, in essence, concerns the conclusions that these researchers arrived at with respect to

[218] See, for instance: Oh & Kim (2004); Woodman & Luck (2004); and Awh & Jonides (2001).
[219] See Fougnie & Marois (2006, p. 533).

attention since, as mentioned above, there is a serious worry that details about the functioning of memory may be confounding the results on attention that such studies would appear to support.

Consider the fact that each of the theories of attention summarized in the sub-sections of section 3.1 relied substantially upon experimental designs that would track attention through monitoring the participant's ability to either shadow, perform from instructions, or later recall and report what he or she could of the presented experimental stimuli. In every instance, regardless of how short the interval between stimulus exposure and response, the participant, in order to be successful, would be forced not only to attend to various stimuli, but to also hold that information in memory for at least some amount of time prior to generating a response. In support of this view, consider the claim made by Prinz (2000) that, if a perceptual input is not in some way encoded into memory, "it cannot be used to select purposeful responses" (p. 252). In other words, perception alone is not enough to enable an individual to deliberately react to stimuli. In cases where participants are required to recall and report upon their experience after an attention task (or a divided attention task), there is scarcely a doubt that memory is involved in the process. Yet, even with respect to auditory shadowing tasks, where the response follows the presentation of the stimulus much more rapidly, there must remain an impression of the words heard upon short-term memory in order for the participant to be capable of faithfully replicating the heard speech. Likewise for performance tasks such as sight reading from a sheet in order to play a piece of piano music. In this case, participants must make use of long-term memory and draw upon stored prior learning in order to enable them to decipher the meaning of the symbols on the sheet before being able to accurately re-produce the piece of music that is symbolically represented before them.

Much more could be said about the relation between attention and memory, but the above ought to be enough to raise a serious concern about the 'purity' of the findings on attention since, in experimental practice, both attention and memory have often been studied without an eye for keeping the two distinct from one another. In fact, as will be seen in the next section, the same worry about the conflation of memory and attention plays a part in a different problem with attention research. The next issue to be considered has its source in a particular and influential multi-component model of memory developed by Baddeley and Hitch in 1974. The problem that will be considered in the following section has to do with a component of their model known as the 'central executive' and its possible role in the control of behaviour.

3.3 The Central Executive

As mentioned at the end of the previous section, one of the early and highly influential models of working-memory was advanced by Baddeley and Hitch in 1974.[220] Far from developing a theory of working-memory that treated memory and attention as separate and distinct processes, Baddeley and Hitch actually incorporated an attentional component within their model under the label of 'the central executive'.[221] According to them, working-memory consisted of three main component systems[222]: 1- The phonological loop (a short-term memory store able to hold and manipulate verbal content); 2- The visuo-spatial sketchpad (a short-term

[220] Their model was essentially a multi-component replacement for an earlier understanding of short-term memory as single unified system which had begun to look inadequate. See: Baddeley (1992, p. 556).

[221] Indeed, Baddeley (1981) goes so far as to say that, "it seems likely that any adequate model of WM [working-memory] will also have to be a model of attention" (p. 22).

[222] However, Baddeley would later propose a fourth component system known as the 'episodic buffer'. See: Baddeley (2002a, p. 91-94).

memory store that could hold and manipulate visual and spatial content); and, 3- The central executive (an attentional oversight system responsible, in general, for regulating inputs and outputs with respect to the two memory stores). The first two of these components are often referred to as 'slave systems' (i.e. systems that operate primarily in an automatic manner) that are under the control of the central executive. The executive system, on the other hand, is characterized as overseeing, coordinating, and constraining the activity of both of these slave systems.

Originally, the bulk of the research pursued by Baddeley and his colleagues was directed upon the phonological loop and the visuo-spatial sketchpad since, according to Baddeley, "it seemed better to concentrate efforts on the more tractable problems of the two slave systems" (1996, p. 5-6). This early avoidance with respect to clarifying and analyzing the central executive component of the model led to charges that it was nothing more than a contemporary title for a homunculus.[223] In other words, the 'central executive' label seemed merely a cover for all that remained to be explained by the model. The worry, in essence, is echoed by Attneave: "if all the responsibility for perception and action is attributed to a homunculus, explaining his behaviour poses exactly the same problem as explaining that of the whole organism, and we have got nowhere" (1960, p. 777). A mature cognitive science is unlikely to tolerate such an idea for long (if at all). But, to be fair, Baddeley openly acknowledged that the central executive played such a role in the beginning; however, he would go on to argue that positing such a thing as a central executive—even though it remained largely unexplained at the outset—served the pragmatic purpose of encouraging focused research into the two slave systems while maintaining an overall cohesive model of working-memory.[224] Indeed, for him, the 'homunculus' that was the central

[223] In this context, the homunculus is essentially a metaphorical 'little person in the head'.
[224] See: Baddeley (1996, p. 8).

executive component was "merely serving a holding function" (2002b, p. 247). Moreover, in time, Baddeley would again take up the problem of elaborating the functions of the central executive and gradually move beyond treating it as a place holder for all that remained unexplained about working-memory.[225]

Baddeley was initially inspired to flesh-out the notion of the central executive in terms of the 'supervisory attentional system' (SAS) model advanced by Norman and Shallice (1980). On their model, "deliberate attention exerts itself indirectly through its effect on activation values" (1986, p. 5). This means that attention need not be directly involved in selection processes. Rather, they maintain that it is through the indirect biasing of neuronal activations that attention mediates the behavioural schemas that end up being adopted by the individual.[226] Early attraction to and influence by the SAS model aside, Baddeley would later propose to analyze the central executive in terms of what he considered to be its four basic functional dimensions.[227] These dimensions, according to Baddeley, included: 1- Focusing attention; 2- Dividing attention; 3- Switching attention; and 4-Interfacing between the two slave systems and long term memory.[228] The study of these four dimensions would be undertaken by way of performing dual-task experiments with frontal lobe patients, Alzheimer's disease patients, and comparative analyses between multiple age groups (among other strategies).[229]

What is of greatest importance to our concerns is the fact that Baddeley's notion of the central executive was (for some time) entirely cashed-out in terms of attentional processes. The

[225] See: Baddeley (2002a&b & 2007).

[226] There are many other important component features of the SAS that I will not go into here. For more detail, see Norman and Shallice (1986).

[227] Although the argument has been made that the notion of a central executive is, by definition, a unitary system (see: Kimberg, D'Esposito & Farah (1997)), Baddeley (2007) maintains that it may nevertheless be "fractionated into subcomponents" (p. 119).

[228] However, he would also consider this last dimension to be a distinct additional component of his model of working-memory dubbed the 'episodic buffer'. See for example: Baddeley (2002a, p. 91).

[229] See Baddeley (2007, chap. 7) for a modest summary of the results of these labours.

worry isn't just that his findings with respect to the central executive functions might be muddled by their involvement as parts of a system of working memory—this problem with attention research has already been identified in the previous section—but rather, the problem is that defining the central executive in exclusively attentional terms fails to capture (arguably) the most pertinent feature of our capacity for *centralized behaviour control*; namely, the volitional and intentional nature of personal agency.[230] In other words, Baddeley's central executive functions, on a second look, appear to be operating more as a middle-management team. That is to say, although the attentional processes that Baddeley identifies may play a role in supporting or constraining the slave systems of memory in some way, they nevertheless do not appear to be the systems that have the final word with respect to the behaviours deliberately or willingly adopted by the individual. And this is surprising since, in one of his earlier papers, Baddeley acknowledged that, "an adequate theory of the Central Executive would probably include not only a specification of its method of manipulating control processes…but would also require an understanding of selective attention and probably of the role and function of consciousness" (1981, p. 21). Unfortunately, many of his later attempts to explain the central executive would seem to have forgotten this earlier recognition of the importance of the role of consciousness. Indeed, in a 2003 paper, he admits to making the "simplifying assumption that the executive was a purely attentional system" (p. 835). However, to be fair, in his most recent work, Baddeley (2007) again returned to consider the role of consciousness in the experience and production of deliberate action.[231] And although he presented some important and interesting research in support of his latest view that consciousness is a crucial component of both working-memory

[230] As we saw in section 3.1.5, Neumann (1987, p. 375) was cautious to avoid this issue as well.
[231] Here he draws extensively upon the global workspace hypothesis of Baars (2002).

and human action in general, how conscious intention or volition operates upon or interacts with other so called executive systems and sub-systems remains underspecified in his analysis.

It ought to be acknowledged that the problem of the avoidance of the notions of intentionality and volition is not restricted to those who would adopt a Baddeleyan model of working-memory or attention.[232] Indeed, the avoidance of these ideas seems to be ubiquitous in the attention literature and research. Many researchers, it would appear, are comfortable enough with making references to executive systems—at least when doing so can't seem to be avoided—since, the functions of these executive systems, after all, may eventually be explained in terms of neural mechanisms. However, when it comes to the issue of explaining how these executive systems relate to or embody the intentions of the individual or what is willed by him or her—or even, how it is that intentions or volitions are psychologically represented at the level of mechanism—attention theorists have for the most part remained silent. To be fair, intentionality would appear to be an exceedingly challenging psychological element to quantify and to measure accurately—not least of which for the fact that it may be inextricably bound up with any number of different psychological contents—and thus, it is perhaps not very surprizing that researchers should aim to steer free from such concerns (at least for the time being). However, as the world outside of the psychological laboratory seems to operate, people tend to be more concerned with whether or not a given *person is intentionally or willfully responsible for his or her actions* than they are with whether or not some supposed executive system was involved in constraining that person's memory or attention for a given period of time that the person was acting. Therefore,

[232] The Baddeley & Hitch (1974) model of working-memory (later developed primarily by Baddeley) serves merely as one of the more salient examples of the neglect of the ideas of intentionality and volition in the sphere of psychological research on attention in general. This neglect is made more apparent for the fact that their working-memory model contains a component that is explicitly characterized as responsible for executing the executive control functions over an individual's behaviour and awareness and yet, how the individual's intentions or volitions are to play a role in that process remains almost entirely unclear.

the broader questions about how attentional systems interact with personal intentionality and volition, as well as what role such factors play in social accountability, remain on the horizon for any account of human behaviour that truly aims at thoroughness.

Attention theorists (and psychological researchers in general), will often refer to 'action-plans', 'behavioural goals', 'schemas' or the equivalent in the place of speaking about the intentionality of agents since, even though notions such as action-plans and behavioural goals may appear to be taken from the same lexicon as intentionality, they are at least readily specifiable whereas the nature and operation of intentionality appears to remain at least somewhat more elusive. I want to be clear that my highlighting of the fact that attention researchers seem to avoid the notions of intentionality and volition is not an indictment of the field. I am sympathetic to the challenges faced by these researchers in attempting to understand and explain the complex functioning of attentional processes in humans (and sometimes in other primates). And explaining something like intentionality, it may be argued, is beyond the scope of projects focused more exclusively upon understanding the basic processes of attention. Fair enough. However, insofar as attention is—at least in some instances—something that is fundamentally controlled by persons, it seems reasonable to anticipate that an explanation of how intentions and volitions not only factor into attentional processes, but how they operate on their own ought to be forthcoming. The disconnect between talk of executive systems, and action-plans and goals, on the one hand, and the intentions or will of an agent or person on the other, is not just a problem for those working on attention.[233] Indeed, it is particularly problematic for a theory of autonomy that looks to draw upon attention research in order to guard against automaticity, since autonomy is explicitly concerned with a person's conscious intentions and

[233] The tension between these two approaches appears to be just one more example of the opposition of causal explanation and teleological explanation.

volitions and an understanding of attention that ignores these notions fails to fully integrate findings on attention with personal agency. In the next section, we will at last consider what role (if any), attention research might play in bolstering a theory of autonomy that seeks to restrict instances of automaticity from interfering with autonomous actions, and further, to bar them from counting as in any way expressive of an individual's autonomy.

3.4 Attention and Autonomy

After having examined the attention literature, it would appear that researchers in this area have been primarily concerned with the constraints upon or limitations of attention (especially as it is examined in the condition of being divided between tasks). Because of this overriding focus upon limitation, it would appear that one of the primary contributions of attention research to our theory of autonomy is to reveal the general thresholds at which attention typically appears to break down for the average person.[234] Beyond this important contribution of helping us to establish the limits of what we can reasonably expect an individual to be capable of attending to, research on conscious attending also continues to support the idea that active attention is to be counted as distinct from and contrary to instances of automaticity which can (at most) only accommodate forms of non-conscious, semi-passive registering of some environmental stimuli.[235]

[234] The details of which (too numerous to mention here) may be gathered by looking at the results of the various empirical studies referenced in this chapter. Importantly, this research is also ongoing and so we may expect our understanding of human potentials in this area to be undergoing perpetual revision and refinement.

[235] I use the term 'semi-passive' here because, as we saw in section 2.1, automatic processes may nevertheless be active enough to respond to changing environmental conditions and yet these processes unfold entirely below the threshold of conscious apprehension or choice.

However, when it comes to the structural outlines of the models provided, none of these, it would seem, are able to provide exactly what we are looking for. For instance, whether or not attention is the result of a structural bottleneck (Broadbent), or contains an intensive element (Kahneman), or consists of multiple resource pools (Wickens), or is constrained by selection processes rather than capacity limitations (selection-for-action view), leaves us with little certainty about how to impede instances of automaticity from interfering with otherwise autonomous actions.[236] That is not to say that attention research has provided us with no insight into this question. Indeed, with respect to the general gradual onset of automaticity, researchers have often been careful enough to recognize when some task within a dual-task experimental protocol has been rendered less demanding upon attention due to the automatization of behaviour resulting from habituation to the requirements of the task. This finding, with respect to the impact of behavioural repetition and the consistent pairing of stimulus and response, sheds some light upon the typical conditions of instances of automaticity becoming manifest. But it falls short of telling us how exactly it might be that actions to which one pays attention could be protected from lapsing into automation, or how attention might be used to impede or halt the occurrence of automaticity altogether.[237] Nevertheless, as suggested above, the individual experiments comprised in this vast area of empirical study do enlighten us with respect to the limits of our abilities to attend—and where it is generally found that human beings are incapable of attending to something (for instance, as a result of one's attention being already divided between other tasks or due to a very low level of alertness), we can assume that, under such

[236] Nevertheless, the selection-for-action view appears to have the most in common with the model of autonomy being developed in this dissertation—and for that reason it is the most appealing of the theoretical frameworks described above to draw from—since both are centrally concerned with developing an understanding of human action and both take a functionalist approach.

[237] Indeed, researchers have, for the most part, simply treated automaticity as something to contradistinguish their views of attention from (e.g. Baddeley 1990, p. 125). And, as far as I am aware, there has been no sustained empirical effort to determine how attention may be used to counteract automaticity.

conditions, a person's actions (or lack thereof) with respect to that thing are incapable of being autonomous. And the same applies to the earlier question about the relationship between raw desires and reflective desires and volitions mentioned at the outset of this chapter. Without the ability to attend to one's raw desires, a conscious and reflective coherence cannot obtain. Indeed, it is implied by the very term 'reflective' itself that such desires and volitions must *reflect* the content of the raw desire in order to cohere with it and produce an autonomous action. In other words, it is impossible to reflectively consider—or mirror or oppose—some raw desire without in some way being able to attend to what that raw desire is.

Although the contributions of attention research noted in the previous paragraph may appear to be somewhat limited, the above is not to suggest that attention does not play a more significant role in the maintenance of autonomous actions. As mentioned at the outset of this chapter, automaticity is in part characterized by a lack of conscious attention to one's present behaviours. So not only is it the capacity to attend to what one is doing, but also actually deliberately attending to one's actions that is clearly of central importance to acting autonomously. Moreover, the attention research has reminded us of the great complexity of human action (e.g. by way of selection problems identified in section 3.1.5) and the many potential systems (e.g. the components of working-memory, or the possible multiple resource stores, or selection mechanisms, *et cetera*) involved in the overall process of attentively acting and responding to one's environment. However, merely attending to what it is that one is doing still does not appear to be enough to provide the kind of support that is needed by the model of autonomy being developed in this dissertation. Indeed, people often describe having experiences wherein they are fully aware of what is taking place within their immediate environment (including their own physical reactions and movements). And yet, often times, they will report

having been unable to react or behave in the desired ways. For example: an individual may slip and take a tumble down a flight of stairs while all along being conscious of what is happening after slipping the first step. Indeed, this individual might even recognize different points at which reaching out for the railing or to extending a leg would be optimal in order to halt the tumbling and yet fail to act on these observations. Such instances are not failures to attend to one's environment (at least, in our example, not after the initial slip), but rather, they are more precisely failures to control one's movements effectively in response to what one is made aware of by way of attention. These sorts of failures highlight a second essential characteristic of automaticity, which is a failure to consciously control one's bodily movements. This type of failure appears to emerge by way of a disconnect between processes of conscious awareness and attention on the one hand, and processes of effective motor functioning and control on the other.

In order to address this latter problem we must look beyond mere attention for a possible answer. Because the problem appears to be characterized by an inability to integrate and coordinate one's conscious contents with one's desired action, we will next turn to research on the unity of consciousness in an attempt to identify just what is required to keep both one's conscious perceptions and desires as well as one's intended behaviours working effectively and in synchrony.

Chapter 4

> No one ever had a simple sensation by itself. Consciousness, from our natal day, is of a teeming multiplicity of objects and relations, and what we call simple sensations are results of discriminative attention, pushed often to a very high degree.
>
> -William James

> …some preliminary clarifications concerning what I mean by 'experience' and 'consciousness' are in order. There is of course a limit on what can be said on this topic: if you do not know what it is like to have experience, words will not help, and there is probably no 'you' there to find out. But although 'consciousness' and 'experience' are to some extent primitive notions, they are as hotly contested as any in philosophy. The literature is full of distinctions between different types of consciousness, theories about what can and cannot be said about consciousness, and the relationship between consciousness and the physical world.
>
> -Barry Dainton

4.0 Introduction

In the previous chapter, we looked at several of the major theoretical views on attention and some of their supporting research. This focus upon attention was motivated in part by the fact that attention to one's behaviours is commonly treated as contrary to automaticity and, as suggested at the end of the first chapter, automaticity is a problem for an adequate theory of autonomy. So it was deemed important to figure out just what role (if any) attention might play in buffering a robust theory of autonomy from the worries associated with automatized behaviours. Although it was determined that attention is central to acting autonomously in

general[238] (i.e. since it remains doubtful on any account that behaviours which are outside of one's attentional purview could be actively 'self-governed'), it was also seen that attention alone was not sufficient to protect autonomy from the threats generated by automaticity. And this is because the mere ability to attend to something is not yet enough to ensure that what one is *aware of* (in virtue of attending to it) coheres in the proper sorts of ways with what one is in fact *doing* (or even with what one simply *intends* to be doing).[239] Thus, in order to guard against the possibility of an agent's otherwise autonomous course of action becoming instead one or more protracted episodes of automaticity, we are going to need to look beyond mere attentiveness for an answer. Indeed, at the end of the previous chapter, we saw that it wasn't just the absence of the agent's attentive concentration upon what she was doing that was central to automaticity, but also that when behaving automatically, she likewise fails to actively consciously *control* her behaviour as well.

The above mentioned lack of coherence between, on the one hand, the agent's conscious attention, intentions, and desires, and on the other, the conscientious control over her own bodily movements is a fundamental characteristic of automaticity that alerts us to what we might need in addition to attention in order to reinforce our view of autonomy. Indeed, if automaticity is characterized not just in terms of a lack of attention to what one is doing, but also in terms of a certain lack of coherence and control between one's conscious, willful mental life and one's behaviours, then examining the research on what is involved in having a unified and coherent experience of the world and one's volitional movements within it would appear to be a fruitful place to search for its antidote. It is for this reason that a substantial portion of this fourth and

[238] With respect to the model of autonomy being advanced and developed in this dissertation, attention was also seen to play a significant role in the coherence between one's raw and reflective desires and volitions.

[239] Take, for example, the phenomenon known as "alien hand syndrome" wherein one may attend to all sort of things that one's limb is doing without being at all capable of intervening in those motions. See also section 3.4 in the last chapter for a similar example of attention without action control that does not rely upon a rare disorder.

final chapter will be devoted to gleaning some useful insights from research on the unity of consciousness in the hopes that we may draw upon some of the relevant findings therein to help us develop a more robust account of autonomy—ideally, one that may withstand the problem of automaticity.

Before we examine the relevant work on the unity of consciousness however, we would do well to say a few words on just what consciousness itself is—or at least, the way in which we will be treating it herein—in order to better understand why and how having a unified consciousness of a particular sort is crucial to establishing a form of autonomous agency sturdy enough to ward off automaticity. It is to these preliminary remarks about consciousness that we turn next.

4.1 Consciousness: A Brief Sketch

As we saw with respect to both of the previously treated concepts of autonomy and attention, there is likewise, a multitude of competing (and in some cases complimentary) understandings of consciousness on offer. Indeed, as suggested in the excerpt by Dainton at the outset of this chapter, there are numerous types of consciousness as well as theoretical views about how to understand consciousness and its relation to the physical world generally. I do not propose to advance anything like a definitive account of consciousness here. Such an attempt would be far too ambitious for a single chapter (let alone a subsection) and it would also divert focus away from the objective of this dissertation, which is to develop a theory of autonomy robust enough to deal with automaticity. Indeed, for our purposes, some basic distinctions will

suffice as an entry point into a specific sub-field of consciousness research that bears more directly upon our immediate concerns; namely, an understanding of the unity of consciousness that can be of service to our project.

Allow me to begin with an incomplete description of my own conscious experience at this very moment:

> I am sitting at my desk shifting between looking at the keys (some of which I am pressing and releasing in a deliberate sequence) and the screen of my laptop as well as the words that I have written and that I am currently writing. I see various other objects atop the desk (books, papers, pencils, *et cetera*), and the sand, orange, and red bricked wall behind it that I am facing. Toward the periphery of my line of vision I see other features of and various coloured objects in my room but in a more attenuated or increasingly 'blurry' way approaching the edges. I smell the incense that I lit to help to keep me a little more alert. I taste the honeyed Korean red ginseng slice in my mouth that is intended to serve the same purpose as the incense. I hear the sounds made by the fan in my humidifier coming from behind me. I feel a general stiffness in my neck and other muscles as well as the position and subtle movements of my limbs and body as a whole. I am occasionally distracted by apparently random thoughts and memories—some partially related to my present concerns while others are less so. And I am pervaded by a slight sense of anxiousness to get this description finished quickly since, a few moments ago, I was informed that a renovations crew will be arriving at my apartment shortly to complete some work and I believe that once they are here, it will quickly get too loud and distracting for me to get any work done at my current location.

This account contains a number of different conscious elements that together make up a recent portion of my experiential life. It mentions, at least in part, what I had consciously *perceived* by way of each of my five senses and via proprioception. The account also spoke to the *sense of personal agency* that I had and that was displayed by the claim to authorship of the thoughts entertained and writing being done as well as the *intention* behind my chewing the ginseng and having lit a stick of incense. It further made reference to intruding *thoughts and memories* as well as *beliefs* and a slight *mood* of anxiousness that was induced by my *understanding* some previously received information. Each of the above italicized components may be considered different types of consciousness, and although several different types were mentioned, they do not by any means exhaust the list of the possible sorts of experiences that one

might have. Nevertheless, what each of the above elements has in common is that they were collectively the components of my *phenomenal* consciousness. That is to say, each of these elements, for at least some stretch of time, contributed to 'what it was like' to be the creature that I am.[240] Another way of putting things would be to say that any given conscious creature has its own point of view and the account just given amounts to my own personal experiential perspective for a brief portion of time. This phenomenal characterization[241] of consciousness will be central to our understanding of what follows.[242]

Another point that might be drawn from the description of my recent phenomenal life is that each of the elements mentioned in the previous paragraph were concerned with one or more 'objects' (i.e. the *incense* that I could smell, the *ginseng* that I could taste, the *intention that* I use both to increase and maintain alertness, *et cetera*). These sorts of 'objects' may be considered the *contents* of consciousness. But there also appears to be a sense in which the underlying background consciousness (upon which such contents make their impressions) itself may be treated as different.[243] For instance, I may see the object 'laptop' in front of me, but it might appear very different to me depending on whether I am in a normal wakeful 'state'[244], or drowsy,

[240] Similarly to Bayne (2010, p. 6) and Block (1997, p. 380), I adopt a 'liberal' view of phenomenal consciousness that includes things like thoughts, desires, beliefs, intentions, emotions, and understandings.

[241] Along with Block (1995, p. 230), I am not confident that I could provide any non-circular definition of phenomenal consciousness, nor am I troubled by that fact.

[242] In general, I will treat the terms 'phenomenal', 'experiential', 'aware' and 'conscious' synonymously.

[243] And this difference can make a difference by potentially restricting the selections of certain contents or the extent to which they are available to cognitive and behavioural control. See: Bayne (2010, p. 7-8).

[244] I refer to a 'state' of consciousness reluctantly here (and throughout) since it can be taken to convey something of a static quality that I am uncertain is ever actually true of consciousness (unless perhaps one is taking a 'time-slice' view of it). To speak metaphorically, talk of 'states' evokes an image of a still or stagnant pond—whereas, I think that consciousness, even when calm, focused, or undisturbed is never quite so stable or fixed. Indeed, the flowing stream metaphor of consciousness appears far more apt to me but I won't press this point here. Rather, with respect to consciousness, I will use the term 'state' for linguistic convenience but I caution the reader *not* to take it for granted that I accept any of the above mentioned or other potentially chimeric qualities that might be implied by such use.

or dreaming, or drugged, *et cetera.*[245] And beyond the potential difference in my experience of the particular content 'laptop' there is also something different it is like phenomenally to be in either of those background states in general. This is not to say that one can be more or less phenomenally conscious but only that phenomenal consciousness can take on a different character in these different background states and that these different background states might play a role in what—by way of contents—is available to functional use. For example, I may be conscious of a cup before me but the addition of another cup does not render me any *more* conscious; that is to say, my having more conscious contents does not thereby provide me with a greater amount *of* consciousness. Likewise, one's dreaming about a cup involves no *lesser* degree of consciousness than experiencing a cup in front of one while wide awake; the difference has to do with what the awareness affords one the ability to do in either state. While awake, one's awareness of a cup can be deployed for action control—for instance, one may take the cup and fill it with tea to drink; whereas, in the dream state (unless perhaps one is having a 'lucid' dream), one's awareness of the cup may give rise to no real control with respect to what one may do with it. Again, different sorts of background consciousness can result in different functional and selective constraints upon the contents of consciousness, but no background state of consciousness is any more or less conscious than another. In other words, in my view, being a conscious creature is an all or nothing affair.

Short of providing an actual definition of consciousness (since this task seems impossible to meet without relying upon circularity and synonymous terms), the above provides a basic idea of the sort of understanding of consciousness that will be important for our purposes. It mentions

[245] This point speaks to our understanding of automaticity in an important way: It suggests that being in an automatic state of consciousness may act to constrain which sorts of contents (if any) are either prominent within, or available at all to phenomenal awareness, as well as the extent to which any such contents can be used for the control of action. There will be more on this in what follows.

a number of the ways in which one can be conscious of things, and it calls to attention some of the different sorts of background character that consciousness can take on. However, there is one further highly useful distinction to be drawn with respect to our understanding of consciousness—a distinction that, as we will see, carries through into the coming discussion of the unity of consciousness in a way that will be central to our concerns. The distinction to be considered next was proposed by Ned Block (1995) in an important paper entitled "On a Confusion about a Function of Consciousness".

One of the important ideas that surfaced in the initial characterization of consciousness given above was that certain background states of consciousness might play a functional role with respect to just what phenomenal contents are selected as well as the extent to which any selected content is available to cognitive and behavioural control. This question about the functional profiles of different types of consciousness is central to the mentioned 1995 article by Block (as well as its 1997 revision).[246] There, Block distinguishes between what he calls the *phenomenal* and *access* forms of consciousness. He characterises *phenomenal consciousness* (PC, or P-consciousness), as we have, in terms of the "what it is likeness" of a given totality of experiential contents or properties. Another way of understanding phenomenal consciousness is in terms of the subjective qualities of experience—also variously called 'raw feel' or 'qualia'—that is, the occurrent experiential richness provided by the senses and/or our internal mental imaginings, recollections, beliefs, desires, and intentions. For each of these mental states (or any possible number of them in combination), there is something it is like for a given conscious creature to be in such a state. *Access consciousness* (AC, or A-consciousness), on the other hand, he restricts to the contents of perceptual inputs that can be used to control reasoning and

[246] Although the initial article mentioned here was published in 1995, unless otherwise noted, the citations and references to follow will be drawn from the later 1997 revised version of the document.

behaviour (p. 379). More accurately, Block claims that "A state is A-conscious if it is poised for direct control of thought and action" (p. 382). By his use of the term 'poised' he means something "intermediate between actual use in reasoning, and so forth, and mere availability for use" (p. 384). To clarify further: by 'intermediate between' actual use and mere availability, I take it that he means to portray access conscious states as 'ready at hand' for deliberate deployment rather than representing something closer to dispositions or bare inclinations or propensities. Indeed, Block's characterization of what AC affords us is essentially given in terms of more active cognitive operations like retrieval, reportability, and the rational control of behaviour.[247] That is to say, it essentially concerns the functional roles of various mental states. He goes on to claim that although there may be no actual cases wherein what one may consciously access is separate from one's phenomenal experience, it nevertheless appears to be the case that PC and AC are at least conceptually distinct (p. 386). And this conceptual distinction appears both important and useful since, for example, we may know what information is available to some other person to report or employ for behavioural guidance without knowing what it is like to be that person.[248] Moreover, the noted conceptual distinctness appears to be supported by the fact that, as Block maintains, AC is a 'functional notion' and that it plays an essentially informational or representational role in reasoning and behaviour control (by way of its causal relations to other representations[249]), whereas he claims that PC "is not a functional notion" (p. 383). Despite their apparent conceptual distinctness, Block suggests that "A-consciousness and P-consciousness are almost always present or absent together" (p. 401), and

[247] There will be more on just what Block seems to want to convey by his particular use of the notions of 'poise' and 'rational control' to follow.

[248] Recall Frankfurt's physician/psychotherapist—he knows what drives his patients to take the drug but desires to feel the pull of the first person craving that they experience in order to better assist them. In other words, he knows about what they draw upon in their states of access consciousness but not what it is like to be in the particular states of craving that they experience (i.e. the states that they phenomenally endure).

[249] See his revised account (1997, p. 384).

that they also interact.[250] For example, when one has the phenomenal impression of, say, a yellow banana peel, that individual is usually capable of reporting on this phenomenal state and using it to guide behaviour (say, to avoid stepping on the peel and slipping). It is perhaps the frequent (if not virtually ubiquitous) empirical co-occurrence of AC and PC as well as their interaction that tends to lead to the frequent conflation of these two distinct understandings of consciousness. In any event, Block's central contention in the article is that an imprudent conflation of these two conceptually separate notions of consciousness has led to a number of confusions and erroneous claims in the research and literature dedicated to the topic. However, I will not comment any further upon the broader value of his proposed distinction to clearing up those confusions here.[251] Instead, I want to consider one more important point about Block's construal of AC before moving on to consider the way in which his distinction applies to and informs our concerns.

One of the primary examples that Block draws upon in order to flesh-out his descriptions of both PC and AC (and their conflation/confusion) has to do with the rare medical condition known as blindsight. Block makes use of both real data derived from blindsight studies and invents extended hypothetical cases in order to press the example into a number of different services. The point of greatest relevance for us, however, is that he uses the standard case of blindsight as a backdrop against which to define the sense of 'rational guidance' afforded by AC. This is meant to add a little more clarity to his particular use of the term 'poised' when speaking about AC. Consider, as background, that in normal blindsight studies, subjects with damage to

[250] One example that Block provides of this interaction has to do with the contrast between the figure and ground of one's experience; here, changing one's perceptual access from figure to ground or *vice-versa* can affect one's phenomenal state as well.

[251] The interested reader may turn to the justifications of the distinction provided by Block himself in both the article's 1995 version which is followed by a number of critiques along with a final response by Block, as well as the 1997 revised version.

their primary visual cortex commonly have 'blind' sections within their visual fields of which they claim to receive no visual information or impressions; and yet, they are nevertheless capable of 'guessing' correctly and reliably with respect to certain features of visual stimuli that are presented only within the 'blind' areas of their visual fields.[252] It is this sort of 'guessing' (regardless of how reliable it may be) that Block wants to rule out when it to comes AC. For Block, such an ability to guess reliably does not amount to representations that are 'poised' for the rational control of thought and behaviour. Indeed, supposing that an individual with blindsight is presented with an 'X' in a blind portion of the visual field, Block claims, "The blindsight patient…has no X-representing A-conscious content, because although the information that there is an X affects his 'guess,' it is not available as a premise in reasoning (until he has the quite distinct state of hearing and believing his own guess), or for rational control of action or speech" (1997, p. 385). Thus, according to Block, when information is only available to an individual by way of being summoned by specific *external* prompts—the absence of which results in a content's showing no influence upon the individual's thought or behaviour—it should not be considered part of AC.[253] Instead, it is only when a given conscious content is ready at hand to *directly* impact thought and behavioural control that it forms part of a creature's AC.[254]

[252] See Block (1997, p. 375) for greater detail and suggested readings.

[253] Indeed, such contents would appear to be closer to inert subconscious impressions.

[254] The astute reader will notice that the previous quote says something a little stronger than this still; namely, it claims that a given content must be available as a 'premise in reasoning' in order to count as part of AC. I'm not certain that I want to follow Block quite that far. It seems to me that to insist that any access conscious content ought to always be available as a 'premise in reasoning' might tether consciousness a little too tightly for comfort to language (i.e. it might render AC too rationalistic). Presumably, other ostensibly conscious creatures that lack language may still deploy some forms of, for instance, conscious cognitive retrieval, or behavioural control or adaptation in response to perceptually experienced stimuli. Instead of requiring that a conscious content be available as a premise in reasoning in order to count as AC, I think that the idea of a given content being 'ready at hand' to 'directly' impact thought and behaviour control is enough to distinguish the sort of thing that Block is after from the sort of subconsciously supported guessing that he looks to exclude. And the benefit of refraining from an overly rationalistic construal is that we may then preserve the possibility that other non-linguistic creatures could nevertheless possess a sort of AC.

Now that we have a thorough understanding of the distinction between PC and AC—as well as a qualified view of the way in which a given content must be 'poised' for use by AC—we can consider how the distinction maps onto our understanding of instances of automatic behaviour. One thing that is immediately apparent when considering a standard episode of automaticity is that there is a significant *asymmetry* between the reach of AC and PC in such cases. And this difference in the scope of what is available to PC and AC during an episode of automaticity appears to be a hallmark feature of such examples. To consider one such potential manifestation of this asymmetry, let us return to the example of Tom provided in section 2.7. There, we saw that Tom's PC was primarily occupied with the daydream of visiting a tropical location for a vacation while he nevertheless continued to automatically perform the gas tank check that he was tasked with. Now, it might be said that although he performed the behavioural requirements of his job automatically (i.e. without full awareness), there was nevertheless something it was like for him to be both performing such a task and daydreaming simultaneously. Of course, the phenomenal difference between his daydreaming while on the job and, say, on his couch at home, might be lost to Tom since, in either case, he would be so attentionally engrossed in the daydream that other details of his phenomenal setting might remain too peripheral to notice.[255] But Tom's presumed obliviousness to the background settings in which his daydreaming occurs does not discount the potential overall phenomenal difference between the two cases. Instead, for our purposes, the most noteworthy constraint to be recognized as operating upon Tom's consciousness while on the job concerns his AC. Indeed, in the example, while on the job, he only has access to the daydream and *none* of his other behaviours for the duration of the automatic episode. He may be 'going through the motions' of

[255] Here we see that, as is typical in such cases, one's focus of attention is more restricted than one's overall phenomenal awareness.

the physical task that he is responsible for performing while in the state of automaticity, but he is in no direct conscious control of those behaviours. Indeed, even in those cases where an automatic sequence of behaviour is capable of making slight, apparently calculated adjustments to certain spontaneous environmental changes on the fly, these sorts of adjustments are not to be understood as the activity of AC. Rather, such adjustments, although they might be made in conformity with an agent's aims or desires, are not 'ready at hand' in the sense of directly available for deliberate action control *by* the presently conscious agent—or at least, not by that part of the agent that is both present and conscious.[256] This is a subtle but exceedingly important point to be clear on since, at first glance, one might think that a (in some sense degraded) form of AC appears to be responsible for the behaviour performed during episodes of automaticity given that the individual may appear to be behaviourally responsive to certain sudden environmental changes. Now, although such behavioural responsiveness may in fact be observed, the important point to remember about such physical activities—despite their potential sensitivity to environmental changes—is that they are not issuing in a way that is directly connected with the sort of control implied by AC. Rather, they occur below the threshold of presently conscious implementation. To connect the point to Block's example of blindsight, we might imagine that a blindsighted patient is capable of automatically ducking in order to dodge a projectile that is heading in his or her direction even though the person may report having no idea why he or she just ducked (assuming that the projectile remained within the blind portion of the visual field). In such a case, although the agent appears to be acting with deliberate intent (i.e. to avoid being

[256] Recall that, as we saw with respect to the work of Reason (1979) in section 2.4, as well as in the Aarts & Dijksterhuis (2000) study presented in section 2.3.3, often times automatic behaviours will revert to previously dominant behavioural scripts as opposed to an agent's currently willed behaviour. And although these scripts will be responsive to a given particular intentional framework that is attributable to the agent, these are often not in sync with the agent's present will. And even where they are in synchrony with the agent's will, they remain so only by chance and not by authentically active conscious control.

struck by the projectile), the behaviour is not at all responsive to or the result of any occurrently conscious state. Instead, it appears to be a purely mechanical response to certain non-conscious (or at least not AC) perceptual stimuli.

The above example of Tom is consistent with what I think is a common feature of the sort of asymmetry to be found between AC and PC during episodes of automaticity; namely, that they are typically characterized in terms of a restricted degree of AC *vis-à-vis* PC. There are other important things to be said about both AC and PC with respect to our concerns with automaticity but more on that later. The important preliminary points about consciousness in general now covered, we will next take a closer look at just what is involved in having a unified conscious experience.

4.2 Unified Consciousness

Research on the unity of consciousness has garnered significantly more attention over the past several years than it has in previous decades. Indeed, starting with Michael Tye's (2003) "Consciousness and Persons: Unity and Identity", followed by Barry Dainton's (2006) "Stream of Consciousness: Unity and Continuity in Conscious Experience", and most recently, Tim Bayne's (2010) "The Unity of Consciousness", the increasing list of publications clearly marks a growing interest in the topic. However, so far, the majority of the work in this area has been centered on attempting to show *that* consciousness (at the level of the subject) is in fact *fundamentally unified* and how we ought to make sense of that raw datum. For our purposes, however, whether or not consciousness is in fact fundamentally unified in some particular way is

not as pertinent a question to ask as the following two are: 1- In what ways can consciousness be unified? And, 2- Can maintaining any particular form of unified consciousness rule-out the possibility of an otherwise autonomous course of action lapsing into automaticity?[257]

With these latter questions in mind, let us consider some of the proposed candidates for the underlying relation that is responsible for the unity of consciousness[258] in order to determine whether any of these options might help us identify a form of unified consciousness that is resistant to automaticity.[259]

First, and perhaps most naturally, we might consider what is known as *objectual unity*. Objectually unified conscious states are states that are focused upon the same object. For instance, one might see and hear the song of a blue jay in flight. Here, the colour of the blue jay's plumes, its shape, its motion through the air, and the sound of its song are all unified in one's conscious experience of the bird in a straightforward way. Clearly, given that the bird is together both seen and heard in this example, we recognize that an object can be consciously unified across different sensory modalities. However, two experiences can be experiences of the same object and nevertheless not be objectually unified. For instance, perhaps the bird's song is projected—say, due to the echoic properties of the landscape—to originate somewhere other than the precise location of the bird. In that case, the modality specific experiences of the visual characteristics and the sound produced would not be objectually unified. Moreover, one's

[257] The answer to the question of whether or not consciousness is in some sense fundamentally unified at the level of the subject does not concern us since, if it turns out that consciousness is fundamentally unified, then it must be a form of unity that entails automaticity—otherwise automaticity would represent a break in that unity—and would therefore be little use to us. Moreover, if consciousness turns out not to be fundamentally unified, it may nevertheless support forms of unity that could help us to overcome automaticity. For these reasons, I will avoid any argument for or against the fundamentally unified nature of consciousness in this section.

[258] Philosophers have proposed a number of different forms of conscious unity in an attempt to identify which sort of unity might play that sought after fundamental relational role. For instance, they talk of subsumptive unity, subject unity, and gestalt unity among many others. However, many of these types of unity will not be given much if any attention here because they fail to provide us with any reason to think that the particular forms of unity that they isolate can be of any help in ruling out automatic behaviours.

[259] My treatment of the kinds of conscious unity to follow owes much to the work of Bayne & Chalmers (2003), and Bayne (2010).

objectually unified states of consciousness need not even concern actual things in the world; for instance, one might be hallucinating the experience of the blue jay (and its shape, movement, and sound) and although the object 'blue jay' itself (as well as its properties) don't in fact exist, one may still have the objectually unified experience of those properties as real and as belonging to a particular object. Indeed, according to Bayne & Chalmers (2003), "for objectual unity, what matters is that two states are experienced *as* being directed at a common object" (p. 25). Of course, such a form of unity is not limited to only involving two conscious states, but rather, it can include any number of conscious states, so long as each state is centered upon the same object. Imagine, for example, that one is holding a blue jay in one's hands to protect it from a cat that had caught and injured it. Here, one's visual impressions of the bird, its chirping sounds, as well as how its feathers and weight feel to the touch, together with the intention to keep it out of harm's way and perhaps a slight mood of sadness for its having been injured can all count as components of a single objectually unified conscious state.

A different yet related form of unified consciousness is known as *spatial unity*. For a given number of conscious states to count as spatially unified the objects of these states must be represented as parts of a single spatial expanse. Returning to the original example of the blue jay, it is not just the object 'blue jay' but also the visible sky behind it, the landscape below it, one's position in reference to the moving bird and a number of other perceived objects in one's environment that together make up a spatially unified conscious state. One of the characteristic features of this sort of unity is that the content of any two or more represented objects is comparable in terms of the spatial relations that obtain between them. Consider again the example of the blue jay that produces a song at a distance from its precise location due to the echoic properties of the landscape. In that case, from the perspective of spatial unity, both the

object ‘blue jay’ and the bird song belonging to it are spatially unified (despite the fact that they are not objectually unified), because they occur within the same overall space.

Another form of conscious unity that appears relevant to our project is *introspective unity*. This form of unity concerns the experienced connection between reflectively approached internal mentations (i.e. thoughts, feelings, remembrances, *et cetera*). These sorts of unified experiences typically contain both a phenomenal ‘what it is like’ character as well as a certain cognitive accessibility that allows for comparison.[260] A related form of unified consciousness is known as the *unity of focal attention*. This form of unified consciousness has to do with an experience of the many aspects of a single item (or a small set of items[261]) and the access relations such an item bears. These relations may include things like the instrumental value of the item attended to *vis-à-vis* one’s goals or plans, whether or not the item poses a threat, *et cetera*. Focal attention is distinguished from a broad field of consciousness—one containing many items—by narrowing in on a single item (or restricted subset) for engagement. There is a mental state in which there is something (unified) it is like to attend to a single item and be simultaneously cognizant of many of its aspects and import. Focal unity thus brings a number of cognitive resources to bear on a single item (or restricted set).

In terms of a more action centered notion of unity, one might also propose something along the lines of a *means-ends unity* in order to capture the conscious experience of carrying out a behavioural plan in the world in light of the reception of certain perceptually conscious information and the relation of that information to one’s beliefs, desires, and intentions to bring

[260] This kind of introspective unity appears to be present in the Frankfurtian form of reflection. It is also in sympathy with the attentional relation between one’s raw and reflective desires and volitions that was identified in section 3.4.

[261] An ‘item’ is here taken to mean anything of which one may be conscious, including desires or internal mental imagery—thus, the unity of focal attention and introspective unity may overlap, though focal attention outruns introspection.

something about. This sort of unity would appear to at least somewhat capture our sense of the effectiveness of our own personal agency.[262] However, this sort of unity relation is unlikely to be fundamental to conscious unity since one may be fully conscious while nonetheless physically incapacitated.

One might also treat the above mentioned types of conscious unity as components of a more encompassing *representational unity*. According to Bayne, "conscious states are *representationally unified* to the degree that their contents are integrated with each other" (2010, p. 10). And such representational integration can include things like the perceptual properties of physical objects (either individually or in greater numbers), the spatial relations between objects (and their relation to the overall spatial 'field' of one's experience), as well as non-perceptual thoughts. Representational unity is equivalent to and is synonymously referred to as *content unity*. Restated, such a form of unity has to do with the connections between objects of consciousness (or the connections between the experienced properties of a single object). If one experiences an object or content, one will also experience other objects or contents and at least some of these items will be experienced as part of single a group. For example, suppose I am currently conscious of the shoes on my feet and the door in the room. If these items were not unified, I would be incapable of answering comparative questions about the two contents such as their proximity to me or whether or not they are the same shade of brown. That conscious contents like these are to some degree integrated is what allows us to compare and evaluate various members of a single group of contents.

[262] This form of unity is also thoroughly consistent with Bratman's planning model of autonomous agency—a central component of which, you may recall, is that we conceive of our power of agency as taking place across extended periods of time. Means-ends unity should thus be understood to be an essentially diachronic (i.e. temporally extended) form of conscious unity.

In addition to representational or content unity, there are also other forms of conscious unity which are broader in scope. The two of these that are of primary importance to our concerns are *phenomenal unity* and *access unity*. As we have already seen in section 4.1, Block's distinction between PC and AC helped us to highlight a prototypical asymmetry discernable between these two forms of consciousness as they are manifest in cases of automaticity. Recall that Block's characterization of the distinction between PC and AC takes the following form:

> Phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action. (1995, p. 227)

Although the original distinction provided above does not make reference to the notion of unity, we can nevertheless modify Block's original account of this distinction to apply to a conception of consciousness as a unified phenomenon. To this end, we may follow Bayne & Chalmers (2003) in calling these new distinctions 'access unity' and 'phenomenal unity' respectively. According to Bayne & Chalmers, these new distinctions can be understood in the following terms:

> Broadly speaking, two conscious states are *access-unified* when they are jointly accessible: that is, when the subject has access to the contents of both states at once. Two conscious states are *phenomenally unified* when they are jointly experienced: when there is something it is like to be in both states at once. (2003, p. 29)

Of course, neither of these types of unity should be taken to be restricted to only dealing with two conscious states at a time[263]—rather, they may contain any number of a specified set of

[263] It is important to also note that, in the cited article, Bayne & Chalmers restrict their analysis of these two forms of unity to cover only conscious states *at a time* (i.e. synchronically) and not such states *over time* (i.e. diachronically). Nevertheless, I see no reason why these forms of conscious unity could not involve states that persist over time, and so I will, without hesitation, employ their distinctions in the service of developing a

conscious states or even the entirety of one's conscious states at any particular instance. In order to provide a little more detail, with respect to phenomenal unity, let us return to the initial characterization (given in section 4.1) of the brief moment of time in which I described my conscious experience as a collection of various related conscious states while I was involved in writing. In that summary of my conscious states, I mentioned that I could both see, for instance, the multi-coloured brick wall in front of me as well as hear the fan from my humidifier behind me (among many other consciously experienced elements). Now, we might think of these two elements of my experience as representationally unified since, for instance, they can be compared with one another in terms of the sensory modalities by which I was made aware of each, and they can also be compared in terms of their locations with respect to one another (among other things). However, there is a unity to these experiences that is more primitive than the unity of their representational content. Indeed, this more primitive unity is captured, according to Bayne, by "the fact that these two experiences possess a *conjoint experiential character*" (2010, p. 10). My seeing the wall first makes a phenomenal impression in consciousness, my hearing the fan likewise makes another such phenomenal impression, and my seeing the wall *while* hearing the fan amounts to a conjoint phenomenal experience with its very own 'what it's likeness' that is not contained by either of the individual impressions on their own and which also precedes any ability to work with those phenomenal impressions comparatively. Of course, I was phenomenally aware of many other conscious states in that description

---

temporally extended account of autonomous agency. The only significant difference between a synchronic approach to the unity of consciousness and a diachronic one to be found in the literature concerns the notion of transitivity. Most philosophers on this topic seem willing to agree that the conscious states of a synchronically unified totality must be transitive (i.e. if at time *t*, conscious state A is unified with conscious state B, and conscious state B is unified with conscious state C, then, at time *t*, conscious state A is unified with conscious state C). However, whether or not this sort of transitivity obtains diachronically is a little more complicated to workout. See Dainton (2006) for a thorough treatment of the issue.

simultaneously, so this more primitive phenomenal unity extends to include the totality of my phenomenal impressions as had *together* for that moment.

Shifting our focus to access unity, we see that each of the elements of the earlier description of my conscious experience over the briefly detailed period were also access unified. That is to say that, each of the elements of which I was phenomenally conscious were also jointly accessible to me in the sense that I could use those conscious contents, singly or in combination, to guide and control my own thought and behaviour. For instance, rather than writing the description as I had, I could have instead written a poem about the sublime perfumed scent of the incense, or the sweetness of the honeyed Korean red ginseng. I could have also simply entertained the thought that these two elements in combination were much more effective for maintaining alertness than either one of them tends to be on its own. Alternatively, I could have written a complaint to my landlord about how frustrated I was of the short notice left to me regarding the impending renovation crew arrival given that I was in a the middle of working on something requiring a certain freedom from noise and intrusion, or I could have slammed my hands down hard on the desk in anger about how it disrupted my goal to continue writing. What's important when it comes to access unity is not that one in fact accesses a given number of conscious elements in order to perform some behaviour or to carry out a particular thought, but rather, that one *can* directly access any of a number of unified conscious contents in the service of controlling thought or behaviour because those contents are poised or ready at hand for deliberate use. Moreover, as we saw with respect to AC generally, what distinguishes access unity is likewise the causal roles that these interconnected conscious states may occupy within the cognitive system.

Because both phenomenal unity and access unity—though remaining importantly conceptually distinct from one another—are more general in character, they may be used in combination with some of the earlier mentioned forms of conscious unity. So, for instance, one might have an introspective phenomenally unified experience. Such an experience might be characterized by the unified 'what it is likeness' of having some, say, happiness evoking emotional memory. Likewise, the experienced spatial unity that one has of one's environment might be part of a greater access unified conscious state which enables one to navigate the terrain effectively, or which allows one to do so in light of the added intentions contained in one's means-ends unity. Of course, like PC and AC in general, both phenomenal unity and access unity interact with one another as well as typically occur together. So, for the most part, from the perspective of an alert conscious mind, one can typically assume that whatever it is that is phenomenally unified within one's awareness, will also be access unified as well. In other words, the totality of the contents of a phenomenally unified state, are often also the same totality of the contents of a simultaneously occurring access unified state. However, as we saw near the end of the previous section, there appears to be a significant asymmetry between the scope of PC and AC during episodes of automaticity.[264] And this asymmetry is carried over into our consideration of phenomenal unity and access unity as well. That is to say that, during episodes of automaticity, the totality of the contents of one's phenomenally unified conscious state does not appear to be identical to the totality of the contents of one's co-occurring access unified conscious state. Moreover, it is this characteristic asymmetry between one's phenomenal unity and one's access unity during such automatic episodes that I believe reveals a key piece of the puzzle in terms of what is needed by our revised model of autonomy if it is to withstand the

[264] As already mentioned, other background states of consciousness—e.g. hypnosis, sedation, drowsiness, *et cetera*—can produce similar asymmetries between the scope of PC and AC.

threat of automaticity. Indeed, the commonly reduced degree of access unity in comparison to the degree of phenomenal unity in such cases speaks directly to the earlier identified lack of coherence and control between one's presently willful mentations and one's movements. Over the next two sections, I will use the insights gleaned so far about the various forms of conscious unity—drawing primarily upon phenomenal and access unity since these show some clear differences between normal and automatic behaviours—to help reassess several examples of automaticity, as well as build a more resilient account of autonomy.

4.3 The Problem Clarified

Now that we have an idea of the sorts of unified consciousness that may help to render our view of autonomy resistant to the worries associated with automaticity—most importantly, the overarching categories of phenomenal unity and access unity—we may begin to formulate an answer to the problem. However, before attempting to construct a solution to the potential interference of automaticity with an otherwise autonomous course of action, we would be well served by returning to our earlier examples of automatic behaviours in order to rehearse the precise ways in which automaticity undermines autonomy as well as to discover the effects that automaticity may have upon our conscious unity.

In the example of Tom, the gas tank manufacturing line worker originally given in section 2.7 and briefly reconsidered above near the end of section 4.1, we see that there are three central components to what is going on at the level of consciousness. First, it would appear that Tom's phenomenal consciousness is unified in this case. That is to say, there is something it is like for

Tom to be engrossed in a daydream while nevertheless automatically preforming the menial task for which he is employed. As noted, we may contrast this 'what it is likeness' with his having an identical daydream while on his couch at home to show that there is an overall phenomenal difference between such experiences, even though what makes each experience phenomenally distinct from the other may be lost to Tom himself due to his being so singularly attentively focused upon the contents of the daydream. In other words, the phenomenal difference between these two possible background settings in which he may have the same daydream might not make the slightest difference to Tom's thoughts or behaviours since, the way in which either of those dim background settings forms a part of his conscious life is merely phenomenally and nothing more. So far, we have identified two of the central components of Tom's conscious experience: 1- his consciousness of both his daydreaming and his unfolding automatic behaviour in the world (that is to say, the totality of Tom's conscious states) is phenomenally unified; and, 2- his attention is restricted to the contents of his daydreaming alone. The final and most crucial component to notice in this example is that Tom's AC, while unified in terms of the contents of his daydreaming (i.e. with respect to his focal attention), nevertheless contains only a fragment of the contents that are phenomenally unified for him. And it is only those contents that are access unified for Tom that may fall under his occurrent deliberate control. Quite clearly then, the fact that some behaviours may take place automatically means that access consciousness is not *totally* unified in such cases. But let us consider some of the other earlier examples of automatic behaviours (from section 2.7) to establish whether or not this restricted scope of access unity is, as I have suggested, a consistent theme when it comes to automaticity.

Recall the example of the aspiring college basketball player who had spent years honing her skill at performing the layup. In that example, in the middle of a game, she took possession

the ball and automatically performed the layup that she had practiced many times before. As with the example of Tom above, her conscious attention was entirely elsewhere (namely, worrying about her poor performance on a recent test that could lower her GPA and get her kicked off the team). During the course of her automatic movements she unintentionally struck an opposing player with her elbow as she jumped to complete a layup and the blow caused the opposing player to suffer a broken nose. A number of other psychological and emotional background details that don't concern us here were also stipulated. What does concern us is that, because she was not consciously attending to her behaviours as they were unfolding (i.e. because she was in a state of automaticity), she both did not recognize the proximity of the opposing player in her way nor could she have stopped the blow from happening. Again, her bodily movements were disconnected from her occurrently conscious control in such a way as to render those movements divorced from anything like a presently autonomous will. Similarly to the last example, we may presume that her consciousness was phenomenally unified since what it is like for her to automatically perform the layup while being absorbed with her poor test score during the game is different from what it is like for her to be focused upon the identical worry and perform the same set of automated behaviours alone at her home driveway basketball net. Of course this phenomenal difference between the two settings might again be very faint or peripheral for her (that is, if she could recognize it herself at all)[265] since her attention is so concentrated on her fears related to the test that the full extent of the difference between her overall phenomenal background experiences in either instance is likely not the sort of thing that she would be capable of fully appreciating while in such a state. Moreover, we also find here the same structural asymmetry that was noted in the last example between the individual's phenomenal unity and her access unity to the contents of her consciousness (as well as that her attentional focus appears to

[265] Remember that recognition is something afforded by AC and not PC.

have set the boundary of her access unity). A further important point to recall is that, due to the sort of repetitive entrainment required for an instance of automaticity to obtain, what seems to be happening is that an initially deliberate and AC (likely even access unified) performance of some behaviour at some later point becomes activated by a subconscious cognitive sub-system. And although such a cognitive sub-system is capable of mirroring a previously deliberate and attention demanding sequence of behaviours it ultimately does so in an entirely mechanical fashion devoid of any occurrent conscious control or oversight.

Another example of automaticity was given in section 2.7 that was somewhat different in character from the two examples treated above. The example drew from my own personal experience as a young martial arts competitor. In this case, the dissociation that occurred during the state of automaticity took the form of an experience of tunnel vision while my body automatically performed various offensive and defensive movements in the context of a tournament team fighting competition that I was engaged in. During the experience of tunnel vision in which I automatically performed the associated physical combat movements, my field of view narrowed to a small circle surrounded by complete darkness and I could barely tell what was going on in that limited window. Within that state, my awareness of the information normally provided by my other senses was entirely absent and the few images that did appear to me at the center of the tunnel were staggered, incoherent (i.e. insensible), and fleeting. Moreover, I had absolutely no recognition or awareness of any of my bodily movements while in the bizarre dissociated tunnel vision state; nor did I have any recollection of what those movements were after the fact.

With respect to the above example, the case could perhaps be made that my consciousness was not phenomenally unified since all of my normal phenomenal sensory impressions—save for

a fragment of my visual field—were entirely absent. Nevertheless, although my phenomenology may have been severely restricted in this state (as compared to my normal alert conscious perception), it was certainly not (as far as I could perceive of it) fragmented or disunified; that is to say that, I did not undergo two separate or self-contained phenomenal experiences. In other words, although what it was like for me to have the experience seemed to have only contained a fraction of the amount and types of contents that are more common to my regular phenomenal awareness, that restricted set of contents—namely, the darkness of the tunnel and the staggered images arising at its center—nevertheless amounted to a single (i.e. unified) phenomenal state. However, even if the critic grants that my phenomenal experience was unified in the above example, the case might then be made that there was no asymmetry between my phenomenal unity and my access unity in this instance, since both states were significantly constrained to only the contents of the tunnel vision experience. While it is true that both my phenomenally unified and access unified states were apparently restricted to the same limited scope of contents in this case, there remains a sense in which my access unity suffered a still greater impoverishment than my phenomenal unity. That my access unity was more significantly constrained than my phenomenal unity in this case is revealed by the fact that access unity (and AC in general) normally entails the ability *use* perceptual data to affect thought and behaviour—and in this case, it was limited to the mere recollection of only the perception of the tunnel vision, and only after the fact at that. So, although both my phenomenal unity and my access unity were concerned with the same contents, the normal effectiveness of that access unity was additionally hindered in a manner that was not suffered by my phenomenal unity at the time. While it may be debated whether or not this fact is sufficient to establish that an asymmetry between phenomenal unity and access unity still obtains here, I maintain that the extra and

uncommon hindrance to access unity in this case does result in a noteworthy contextual difference between the two forms of unity. It is obvious that the asymmetry identified in the earlier two examples was an asymmetry between the scopes of conscious contents of each distinct form of unity; whereas here, the asymmetry concerns not the scope of the contents, but rather, what is normally afforded by the sort of consciousness in question. Thus, in this example, the asymmetry at issue concerns the normal powers of the types of unified consciousness being compared. Phenomenal unity typically contains a number of 'felt' conscious impressions that are had together, whereas access unity typically allows us to do something with those impressions (or contents) in terms of deploying them in the service of thought or behaviour control. Because my access unity lacked these additional standard powers in this case, the normal symmetry between phenomenal unity and access unity, I maintain, was thrown out of balance.

Another important difference between the tunnel vision example and the other two examples provided earlier is that there was no real error or accident involved in the current case. Recall that the earlier two examples of automaticity resulted in behaviours that each subject would disavow and reject as unintentional and involuntary. Still, there is nothing about my automatic physical performance and its having led me to victory in the present case that I would want to disavow. I entered into the match with the intention of performing to the best of my ability against an opponent that I believed would out-perform me, and my resulting win was nothing that I would reject as undesired or unintended by me. However, as mentioned in section 2.7, there is no denying the fact that from my perspective, as the match was taking place, I was for the most part unaware of what was happening (except for the few unclear images seen through the tunnel which held no meaning to me at the time), and I had no sense of control over my body nor knowledge about what it was doing as the match unfolded. As stated previously,

my body could have done a number of different things while I was in that state: it could have stood frozen with fear, it could have fled, it could have behaved in a purely defensive manner, *et cetera*. No matter what might have happened, my conscious will had no input in the matter once I became dissociated as a result of the tunnel vision state. Thus, my automatic behaviours at that time were simply not occurrently self-governed or autonomous. And this state of affairs reveals another important implication of the sort of dissociation that is characteristic of automaticity—namely, that it isn't enough for one's behaviours to by chance simply 'fall in line with' one's earlier desires and intentions, since this might happen entirely outside of the active control of the individual.[266] Instead, what seems needed is both an ability to attend to what it is that one is in fact doing, as well as an introspective access unity that is able to guide one's behaviours in an attentionally responsive way in light of one's occurrently coherent reflective volitions. So it would appear that one's coherent intentions must also be introspectively access unified across time.

The final example of automaticity that we will consider was also the first such example mentioned at the end of the first chapter in section 1.4. There, our central concern was raised in light of Bratman's temporally extended account of autonomous agency. Now, as was mentioned in section 1.4, because Bratman's decision to ground the subject's autonomy in the temporally extended plans and policies that she has was motivated as a response to the regress objection to the earlier Frankfurtian (and Dworkinian) model of autonomy—an objection that we have found another way to address—we will abandon the notion that such plans or policies constitute the autonomous agency of the individual subject. The decision to abandon this component of Bratman's view is also due to the worry that was advanced in section 1.4 that an agent may still

[266] The point made here is reminiscent of the problem of 'deviant causal chains' that was identified with respect to the standard belief and desire account of actions. See: Davidson (1973, p. 78-79), (1978, p. 87) and, Anscombe (1989, p. 378).

become dissociated from her plans and policies when she is in a state of automaticity. What we will keep in focus however, with respect to the Bratmanian view, is the understanding that what supports the individual's autonomous agency are the interconnected psychological ties between one's self and one's memories, one's future oriented intentions and their fulfillment, and the continuities between one's desires and their kin. In addition to the overall coherence of these diachronically dispersed components we will need to add certain other features of consciousness considered so far. But first, let us return to the original example and the worry that it generated.

In the example, we met an individual with the self-governing policy of always standing up for (or coming to the aid of) any other who appears to be in distress. This policy, it was suggested, had long since been entrenched—to the point of having become automatic—by this individual's having always defended other children from the assaults of schoolyard bullies throughout childhood. The policy might have been reflectively adopted in the beginning, perhaps because the individual saw it as the right thing to do. And perhaps such a choice was reinforced by the thanks that would be returned by the victims that were spared from further abuse. In any event, later in our subject's adult life, the policy would again become activated in the context of witnessing a mugging taking place. Unlike the minor risks or dangers attending allowing this policy to become activated in the schoolyard as a child however, as an adult, and in this more serious context, the activation of the policy might come at a far greater cost to the individual. As mentioned, the mugger might have been concealing a firearm and might respond by opening fire upon our would-be helper, perhaps shooting the initial victim first. It goes without saying that, neither of these violent outcomes would be the sort of thing that our agent would invite under normal circumstances. Nevertheless, because the policy was activated automatically, the

behaviours that it engendered—and the consequences that followed directly from them—were outside of our agent's occurrent volitional control.

The most salient difference between the above example and the ones treated just before it is that, in the present case, the agent's attention to a situational trigger actually initiated the automated policy activation; whereas, in the earlier three cases, the agents' conscious attention and control was constrained such that they each did not have access to the automated behaviours as they were unfolding. More precisely, in the earlier examples, an agent would have started out consciously with respect to some course of action, but at some point along the way, key parts of the agent's consciousness simply receded or withdrew from actively monitoring those behaviours and instead allowed non-conscious systems to continue to automatically carry out those behavioural sequences. Along similar lines Bayne notes, "the agent *as such* seems to recede when dealing with representations that are able to drive only a restricted range of consuming systems, and with it our evidence that we are dealing with consciousness may also recede" (2010, p. 104). Our most recent example, on the other hand, reveals a somewhat different kind of automaticity at play. Indeed, in this example, our agent is fully conscious of the fact that another individual is being mugged. In fact, it is our agent's ability to recognize this information that allows the policy to be brought into the service of generating behaviour. However, the activation of the policy itself and the activity that it engenders is not something over which our agent has any occurrently deliberate control.[267] So, whereas the earlier cases are more clearly understood to involve a 'dropping-out' of conscious attention and control, in this case, attention plays a role in the activation of the automatized policy itself.

[267] Recall the following comments by Bargh cited in section 2.1: namely, "...behavioral and cognitive goals can be directly activated by the environment without conscious choice or awareness of the activation" (1997, p. 47); and further, "...evidence demonstrate(s) that action tendencies can be activated and triggered independently and in the absence of the individual's conscious choice or awareness of those causal triggers" (2005, p. 38).

To some, the above state of affairs might make it appear as if consciousness is in fact access unified with respect to the automatic behaviour that follows from the policy activation since: 1- the agent at some earlier point did in fact coherently and reflectively endorse the policy in question; and, 2- the appropriate situational trigger successfully led to the activation of the appropriate policy; while, 3- the agent was consciously attending to that very situation when the activation occurred. However, the activation of the policy in this case was entirely outside of the agent's occurrent conscious control. Indeed, in this example, our agent is no more in control of the activation of the policy that initiates behaviour than a participant in the early 1935 Stroop study (explained in section 2.3.3) is consciously in control of the interference produced by word stimuli upon colour identification. Put differently, the activation of the policy in this case is closer to a reflex than something that answers to the agent's immediate and direct conscious choice. And this is the sort of thing that Block sought to rule-out from counting as AC. Clearly then, this last example does not display total access unity (again, because the agent was not in direct control of the policy's becoming active). Nevertheless, the example does appear to fit with the earlier mentioned notion of means-ends unity since it successfully and coherently connects the previously adopted intention or policy with current perceptually relevant information and behaviour that conforms to the aims of the earlier intention. But this seems more like a reason for rejecting the idea that means-ends unity necessarily underlies autonomy than a reason for accepting it.

The final important point to be made about this last example is that the earlier noted asymmetry between phenomenal unity and access unity appears to be preserved in this example as well. The difference may not appear to be quite as pronounced as it was in the first two examples but it can still be seen to be present. For instance, our agent, in this example, has a

phenomenally unified experience of witnessing the mugging that is taking place as well as what he says and perhaps physically does to try and intervene (among other things). Additionally, although his behaviours might be automatically initiated, they are nevertheless within the scope of his attentional grasp. Interestingly, however, there doesn't seem to be a 'something it is like' for his automatized policy to become active; the agent's phenomenal unity will include phenomenal impressions related to the behaviours that result from such an activation but the activation of that policy itself doesn't appear to have any phenomenal content. In terms of the agent's access unity however, we see that not only is the automatic activation of the policy not under the control of the agent's AC but neither is the speech or behaviour that flow directly from that policy entirely under the control of an access unified consciousness. Even if we were to grant that the agent might have some limited AC control here—assuming, for instance, that he might be able to control whether his intervention remains purely verbal or instead involves physically defending the victim— he is nevertheless not in control of the fact that he intervenes on behalf of the victim (this activity is ensured by the automatic policy activation). Therefore, it is again the case in this example, as it was with the previous one, that the agent's access unity is more severely constrained than his phenomenal unity. And that asymmetry, once again, concerns not the scope of the contents, but rather, what is normally afforded by the sort of consciousness in question.

With each of the above examples reconsidered in light of our understanding of the over-arching phenomenal and access forms of conscious unity (and disunity), we are now well situated to provide a positive account of what is required of a theory of autonomous agency if it is to rule out the possibility of automaticity. In the next section, that positive account will be given.

4.4 The Solution

To begin with, we ought to take into account the pieces of our proposal that have already been provided; namely, those desires and volitions that must be coherent in order to support autonomous agency. Prior to making any changes to our internalist approach to autonomous agency (other than having abandoned the hierarchical language and conceptual framing of the early formulations), the image that we have of what is required for an agent to act autonomously includes a set of coherent raw and reflective desires and volitions that are (more or less) temporally extended. Moreover, for these mentioned reflective volitions to count (i.e. for them to be effective), they need to ensure that the desires with which they cohere are sufficiently powerful to carry the agent "all the way to action" as Frankfurt argued.[268] Also, because some of the desires that an agent might have can take the form of long-term plans and policies, we have to include such forward-looking intentional structures within our overall apparatus of coherent willing and desiring. Finally, we know that because certain cognitive structures, like one's policies, may become automated, that they may generate behaviours that fail to match-up with an agent's occurrent and greater psychological unity (i.e. in terms of the coherence between the agent's present willing and desiring). And that even when such automated behaviours are by chance consistent with the agent's willing and desiring, they are nevertheless outside of the active control of the agent and thus do not amount to examples of genuine self-governance. Moreover, it is not just the danger of automated policies but the simple fact that behaviour can and often does unfold in a manner that is automatic that threatens the agent's ability to behave consistently autonomously.

[268] Of course, put in this imprecise sort of way, such a power is not on its own sufficient to rule-out automatic behaviours.

Now for a first pass at what is additionally required to block a course of autonomous action from succumbing to automaticity. The first thing that appears to be needed in order for any sequence of behaviours to count as autonomous is that the behaviours in question must fall within the agent's attentional purview. Indeed, if an agent is unable to even attend to what it is that she is doing, it is hard to imagine that she could be actively "governing" her own behaviours. However, we are also already aware that mere attention to what it is that one is doing is insufficient on its own to guarantee that one is behaving autonomously (since one can attend to behaviours that are entirely outside of one's intentional control). Clearly then, we need to build a certain type of active control over what it is that one is doing into our model of autonomy. As we have seen in the previous section, one of the ways in which an agent's control over her own behaviour fails during episodes of automaticity has to do with the asymmetry that is standardly found between phenomenal unity and access unity in those cases. During such instances, access unity appears to be consistently under greater constraint (in terms of having fewer contents or a reduction to the normal capacities that it has) than phenomenal unity. Therefore, it makes sense to propose that in order for a sequence of behaviours to count as autonomously self-governed, they must have also flowed—in a continuous manner across time—from an agent's unhindered and *symmetrically unified* phenomenal and access conscious unities. In order to keep things a little more manageable, we will refer to this kind of conscious unity as simply *symmetrical unity*.

One thing that becomes immediately apparent with respect to this new requirement of symmetrical unity is that there is both a broad and a narrow way of conceiving of that unity. Beginning with the broad view, one might propose that the symmetry in question must hold across the totality of the phenomenally unified conscious state and the totality of the co-occurring access unified conscious state in order to support autonomous agency. The narrow

version of the claim for symmetry, on the other hand, might require only that the symmetrical unity obtains with respect to the specific actions undertaken by the individual. Under this view, it matters not whether there may be some discrepancy between the totality of the agent's phenomenally unified conscious state and the totality of her parallel access unified conscious state, so long as those contents which concern our agent's actions in the world are symmetrically unified.

Let us return once more to the example of Tom in order to help us flesh-out the difference between the broad and the narrow theses. The advocate of the broad view might argue that the totality of Tom's phenomenal conscious unity included the sequence of behaviours that he automatically performed, and thus, if Tom's access conscious unity had only extended to include the same totality of contents, then Tom would have had access to those behaviours and he would have also been in complete conscious control of them (and therefore acting autonomously). The advocate of the narrow view, on the other hand, will suggest that complete symmetry between Tom's total phenomenally unified conscious state and his parallel total access unified conscious state goes further than is needed. Instead, the advocate of the narrow view will insist that all that is required in order for Tom to be acting autonomously is that he be in a state of symmetrical unity with respect to the behaviours in which he is engaged.

Given that we have a choice to make between the broad and the narrow theses expressed above, something ought to be said about which approach appears to be preferable. I am persuaded to adopt the narrow view, but allow to me to show why I think it to be the better option by way of an example: In this scenario, let us assume that a driver named Jack has just witnessed an automobile collision at a busy city intersection in which both of the drivers and a cyclist were seriously injured. Jack, believing that the cyclist is the most seriously wounded

deliberately chooses to take the injured cyclist into his car and rush him to the nearest emergency room. As Jack is transferring the cyclist from the scene of the accident into his car the totality of his phenomenally unified consciousness includes the sound of an approaching ambulance siren. However, because Jack's attention is so focused upon getting the wounded cyclist into his car, let us stipulate, he is not access conscious of that approaching siren. Here, I would argue that, despite Jack's reduced degree of access conscious unity in comparison to his total phenomenally unified conscious state, he is nevertheless behaving autonomously—assuming his actions flow from an immediate coherent reflective volition and his attention and AC to what he is doing—with respect to his moving the cyclist into his car in order to get the accident victim to a hospital as quickly as possible. Of course, it is probably a bad idea to attempt to move an injured person oneself (especially if one lacks any sort of medical training or know-how). But it is not Jack's poorly reasoned choice to do so that is under examination here. Indeed, it seems obvious that people frequently make poorly reasoned decisions and nonetheless carry out various actions based on such decisions entirely autonomously. What is at issue here is whether or not Jack's access conscious unity had to be symmetrical to the totality of his phenomenally unified consciousness in that moment in order for him to behave autonomously. I don't think that such a state of affairs is necessary. Of course, the advocate of the broad view could argue that had Jack's co-occurring access unity been symmetrical to his total phenomenal unity at the time, then Jack would have noticed the siren and could have paused and waited for the ambulance to arrive and for the EMS workers to tend to the wounded. While such a state of affairs might have been the more prudent course of action given all of the relevant situational information, I submit that his actions in the first instance are no less autonomous than they are had his symmetrical unity been total (and his behaviour therefor different). And this is because those contents of his

phenomenally unified and co-occurring access unified consciousness were symmetrically unified over the actions he was engaged in performing. Jack's having AC to the sound of the approaching siren would have widened the scope of actions available to him to choose from—for instance, it would have given him an additional reason[269] to wait for help to arrive—but even from within his narrow symmetrically unified state, his actions are fully within his deliberate conscious control. And we can easily imagine Jack himself asserting that his actions with respect to moving the cyclist were entirely autonomous. He may grant that he might have acted differently had he recognized (i.e. had conscious access to) the siren, but that point does nothing to undermine the fact that the actions that he did perform were completely under his willful and direct conscious control.

So far, we have identified two of the central additional components required for an automaticity proof model of autonomy: 1- Attention to what it is that one is in fact doing; and, 2- a narrowly symmetrical conscious unity of what one is doing. But it will be helpful here to say a little more about what exactly the above two points ought to entail. As mentioned in section 4.2, phenomenal unity and access unity (similarly to representational or content unity) are two general, more encompassing ways of thinking about conscious unity. Indeed, it was shown that some of the other more specific forms of unity could be spoken about in combination with the overarching labels of phenomenal or access unity. Thus, one could be said to have introspective phenomenal unity or, spatial access unity for example. And a number of these different and more specific forms of unity also appear important to acting autonomously. First, it seems clear that

[269] This point reminds us of one of the central worries motivating the externalist "responsiveness to reasons" approach to autonomous agency identified all the way back in section 1.1; namely, that having an insufficient handle upon the many reasons there are for behaving in one way or another could be seen to undermine one's autonomy (especially if being so under-informed leads one to perform actions that end up being against one's best interests). However, since we have already rejected such an externalist approach, we won't spend any more time on this point here.

what we earlier referred to as the unity of focal attention plays a significant role in our behaving autonomously. But this form of unity is essentially already incorporated into our model in terms of the requirement that the agent must be attending to what it is that she is doing. It is not just that the agent's attention must be unified upon some object or conscious content but rather, that attention must be focally unified upon those behaviours that the agent is engaged in performing that counts. And that focal attention to what it is that one is doing needs to be unified in order for one's behavioural control (i.e. AC) to be coherent and responsive to the circumstances of the unfolding of those behaviours. Therefore, when it is said that autonomous agency requires that an agent attend to what it is that she is doing, it ought to be assumed that the sort of attention involved is also unified in this particular way.

Another more specific form of unity that is important to our understanding of autonomy is introspective unity. The way in which this form of unity is most closely related to our concerns is that it acts as the bridge between our coherent reflective volitions and our access unified conscious behavioural control. In other words, it is our introspective access unity that not only allows us to control our behavioural movements but which also ensures that those movements are responsive to our greater psychological coherence—i.e. our autonomous willing and intentions.[270] Introspective access unity does this by ensuring that our coherent volitions are directly capable of shaping or guiding our behaviours. Crudely put, it is the glue that holds our coherent intentions and deliberate behaviours together. Without this specific form of unity one might nevertheless consciously control one's behaviours *or* have coherent volitions, but one's controlled behaviour would not be the result of one's coherent volitions. Therefore, we ought to assume that narrowly symmetrical conscious unity encompasses this more specific form of unity.

[270] In fact, to include introspective access unity within our model is essentially to say that one's entire occurrent psychological economy ought to be coherent in order to act autonomously.

There may be other sorts of circumstances where things like object or spatial access unity appear to play a vital role in shaping one's autonomous behaviour as well. However, I think that for the most part, the relevant aspects (i.e. contents) of either object or spatial access unity would already be included within the agent's unity of focal attention upon what it is that she is doing. So I won't look to build these other specific forms of conscious unity into the model being developed here.

To reiterate in a more precise way the key components of our new and improved internalist model of autonomy, we have: 1- A coherent psychological economy that includes sympathetic raw and reflective desires and volitions (which may be enshrined in certain long-term cognitive structures like plans and policies); 2- Unified attention to what it is that one is in fact doing; and, 3- a narrowly symmetrical conscious unity of what one is doing (which includes introspective access unity). All of these components, I believe, are jointly necessary and collectively sufficient to ensure autonomous agency in light of the threat of automaticity. In any sequence of behaviours that one of the above three components fails to obtain, I think we have good reason to doubt that an individual behaved autonomously.

Now that we have provided a positive characterization of the additional conscious elements needed to ensure autonomous agency against the menace of automaticity, we may turn to several challenges to this model to assess whether or not these opposing views warrant any further modifications or deeper structural changes to our proposal.

4.5 Objections and Replies

In the previous section, the example of Jack brought back to our attention a worry that was partly responsible for motivating the externalist "responsiveness to reasons" view of autonomous agency given in section 1.1. This motivating worry revealed that an agent's having a paucity of reasons for acting tends to undermine our confidence in that agent's ability to act autonomously—especially where having a very limited grasp upon the reasons for action might lead to the performance of behaviours that run counter to the individual's best interests. Despite this apparent concern about the number of reasons for action available to an agent, advocates of the responsiveness to reasons approach might nevertheless attempt to challenge our improved internalist model of autonomy advanced in the previous section by claiming that we often do have sufficient reason to treat automatic behaviours as still autonomous. Here, the responsiveness to reasons advocate might argue that it is frequently the case that when an agent behaves in a way that is automatic, her behaviours are nevertheless governed by the reasons for action that have shaped her choices. To return to an earlier example, they might claim that our star basketball player was in fact behaving entirely autonomously when she performed the layup that injured the member of the opposing team. They might insist that her behaviours were consistent with her selected reasons to perform (by any means) to the best of her ability during the game. And, although she may not have intended to injure the opposing player, her automatic performance of the layup was conditioned by and concordant with the reasons she had to play well. Turning to the empirical research, we see that, to a certain extent, the above challenge appears to be borne out by some of the findings on automatic behaviour. Recall the work done by Reason (1979) covered in section 2.4, as well as the Aarts & Dijksterhuis (2000) study presented

in section 2.3.3. In both of those studies, it was seen that automatic behaviours will often revert to previously dominant behavioural scripts—many of which we can assume to have been the result of an agent's prior deliberate and reasoned selections for action (or so the responsiveness to reasons advocate might argue). Indeed, the type of repetition required for any sequence of behaviours to become regular or dominant enough to be activated in an automatic way would seem to defy their having been the result of spontaneous or entirely un-reasoned movements during the drawn-out conditioning phase. Given the apparent supporting evidence just mentioned, advocates of the responsiveness to reasons approach might then argue that we shouldn't have a problem with—at the very least, occasionally—treating such behaviours as the outcomes of an individual's reasons for action or, according to them, her autonomous agency.

While I think that the above challenge deserves to be considered, I nevertheless believe that it fails to seriously undermine our proposal for a number of reasons. First, the argument given that some automated behaviours ought to nevertheless be considered autonomous entirely fails to capture the phenomenology of agency as it unfolds. Moreover, such a view ends up unrealistically simplifying autonomous action and appears to replace its explanation with simple ascription. With respect to its failure to accurately portray the phenomenology of agency, it would seem that by not acknowledging the experiential distinction between actions that are performed (as I would label them) autonomously, and those behaviours that are performed automatically, it misses something essential about the active mental lives of agents as they act deliberately in the world. Indeed, this experienced difference between genuinely autonomous and automatic states is what underlies the agent's acknowledgement of her authorship and personal control of an action or her lack of control and potential disavowal of such behaviours respectively. In other words, such a view amounts to a severely impoverished—or worse, utterly

lacking—account of the richness and involvement of an agent's immediate conscious willfulness with respect to the choices she makes and the actions in which she engages. Furthermore, behaviours performed while in a state of automaticity are often repudiated by the very people having performed them, and they are rejected in this way *because* behaviours performed automatically do not allow for the direct conscious control of action. But if such a lack of control can amount to plausible grounds for the disavowal of behaviour, then that very same lack of control should, at the very least, undermine our confidence that any behaviour produced in such a way could be legitimately considered to be autonomously authorized or regulated. Here, common sense would appear to dictate that if automatic behaviours can legitimately be rejected as involuntary because of the fact that they are automatic and not actively controlled (which it clearly seems they can), then these same sorts of behaviours cannot also be plausibly treated as part of actively 'self-governed' autonomous action. It is this failure to distinguish between occurrent consciously controlled actions and automatic behaviours that amounts to a gross oversimplification of autonomous agency. Rather than explaining the phenomenological difference between the two states, it simply treats any behaviour that can be traced back to some selected reason for action (regardless of, for example, how temporally removed the reason is from the behaviour) as autonomous. In other words, no matter how temporally, cognitively, or otherwise strained the connection between the reasons to act and the behaviours they are supposed to produce, insofar as that connection can be made at all, the responsiveness to reasons view suggests that we can attribute those behaviours to an autonomously reasoned selection; and that this therefore renders the behaviours themselves autonomous. But this approach will often fly in the face of the agent's disavowal of such behaviours as involuntary. That is to say that, as has already been mentioned, although the behavioural sequences in question may have been

conditioned by a particular intentional framework that was at some point in time accepted by and attributable to the autonomous functioning of the agent, these behavioural sequences often do not match-up with the agent's present will. And even where such behaviours are compatible with the agent's will, they remain so only by chance and not by authentically active conscious control. For the above marshalled reasons against treating automatic behaviours as legitimate expressions of an individual's autonomy, I think that the challenge here considered, framed in terms of the responsiveness to reasons based approach, fails to amount to a significant difficulty for our proposal and should therefore be rejected (i.e. independently of the reasons for rejecting the responsiveness to reasons approach in general).

While the previous challenge to our model attempted to include at least some automatic behaviours within the domain of autonomous agency, the next challenge to be considered holds that all actions contain elements of automaticity and thus, one cannot help but behave automatically to some degree. If this is true, it seems impossible to conceive of autonomous agency as it has been modelled in this dissertation. The challenge can be put simply by considering that, for any given autonomous action, there are a number of smaller contributing actions that are not under the direct control of AC, or that are not attended to, or which do not precisely figure into the agent's coherent desiring and willfulness. For instance, let's assume that I autonomously get up from my seat and get a glass of water from the kitchen. That is to say, I have the coherent raw and reflective desire and volition to get up from my seat and go get a glass of water from the kitchen and, that I am attending to what it is that I am doing (in a unified way) while in a fluid narrow symmetrically unified conscious state for the entire course of the action. Now the challenge might be framed along the lines that although the overall or general structure of the action in question may be considered autonomous (on our view), it is nevertheless built-up

from smaller actions, many of which are not performed autonomously; and thus, our account of autonomy must be false. The fact that I lifted myself from the chair using my arms instead of not using them or that I then immediately turned right as opposed to left to get around the chair on my way to the kitchen, or that I walked instead of ran or skipped—and we can begin to see countless other potentially unconsidered fine details that could emerge between the consciousness of the coherent volition and completion of the action—are not only not precisely accounted for, but they may also fail to be the result of direct AC control. In other words, they may be closer to being automatically selected—even though they will likely still be consistent with my will. Let us call this objection the 'problem of precision' (PoP) since it identifies a worry about the more fine grained aspects of our ostensibly autonomous actions.

One noteworthy difference between the previous objection and the current worry is that, with respect to the former, the automatic behaviours in question might not sync with the agent's occurrent will, whereas, with respect to the PoP challenge, it seems the same problem is precluded. But this fact alone doesn't reduce the concern that the problem raises since, even though in the PoP case the more fine grained potentially automatic behaviours, it appears, would likely always be consistent with the agent's greater occurrent will, they appear to remain uncontrolled to a certain extent (at least, this seems to be a genuine possibility) and the noted lack of control was one of the key reasons why we objected to the previous challenge, so it would be inconsistent to deploy the bare compatibility between an agent's behaviour and volition for a contrary purpose here (i.e. we cannot claim that those component behaviours fail to be of concern simply because they may always be consistent with the agent's occurrent willing). So the mere consistency of a given behaviour with an agent's will does not appear to amount to a workable answer here.

It should also be recognized immediately that the PoP challenge seems to present a far more significant worry than the previous one. Indeed, while the previous challenge may have only occasioned a slight modification to our model if it had been successful—*viz*. to allow some automatic behaviours to count as autonomous in certain limited cases—the present objection, if it succeeds, would appear to render our model irredeemable since it would make automaticity inevitable.[271] However, I don't think that we are without recourse against this particular difficulty. In fact, there are a number of rather straightforward responses to this sort of objection. One is to reject the claim that some (or all) of the smaller 'actions' which make-up the larger supposed autonomously willed action are automatic *simply because* they appear to be *indeterminately* drawn from an indefinitely large pool of potential actions. As Davidson (1963) notes with respect to the act of turning on a light:

> If I turned on the light, then I must have done it at a precise moment, in a particular way—every detail is fixed. But it makes no sense to demand that my want be directed to an action performed at any one moment or done in some unique manner. Any one of an indefinitely large number of actions would satisfy the want and can be considered equally eligible as its object. (p. 6)

Following Davidson's lead, we might argue that although those smaller component action details are not identified in our more general account of the autonomously willed action, they nevertheless continue to meet our requisite conditions for autonomy. However, I am somewhat hesitant to entirely get behind this particular line of response.[272] When I consider my having got

[271] Such a problem however does not appear to only afflict the model of autonomy developed here. Instead, it would likely be problematic for any model of autonomy and perhaps even for action in general if successful. Thus, it would appear to entail that, unless this worry can be successfully addressed, any theory of autonomy would fail Dworkin's (1988, p. 7-8) empirical possibility criterion.

[272] One reason for this hesitation is that I'm not sure that we should treat more precise component movements as individual objects for analysis (to do so would seem analogous to adopting the time-slice approach to understanding consciousness which I find dubious). Instead, it seems that our standard evaluations of autonomy

up from my seat to get a glass of water from the kitchen, my immediately turning right as opposed to left to get around the chair does not appear to have been specifically dictated by my guiding coherent volition (which was to get the glass of water). And although, as a component action, it partially satisfies my will, it's not entirely clear whether my more fine grained movements on the way to performing some more general action are as thoroughly or consistently controlled by my AC in the same way that they might be when I say, perform a very simple and direct action like autonomously raising my right arm. Of course, it would appear that one could continue to press the PoP challenge in an ever finer or more subtle way; for instance, by arguing that even the simple action of my raising my right arm meant that my right hand took a certain trajectory through space that was not itself precisely selected or chosen by way of my AC. Taken to such an extreme, we might begin to wonder just how precise our control would have to be in order to count as autonomous in response to this problem. And I think that at the more extreme end of the problem is where we start to recognize the cracks in this particular objection. It seems to me that we simply do not individuate our autonomous actions in such a fine grained manner for ourselves—so again, the current objection seems to not be responsive to our own phenomenology or the level at which we psychologically understand ourselves and our own movements.[273] Moreover, if those smaller component 'actions' are under our unified focal attention and symmetrically unified consciousness because they are components of what it is that we are in the process of (autonomously) doing then I think that though they may not be precisely

---

refer to more general complete actions that are framed and constrained by the reflective volitions of the agent and not some exhaustive account of the specifics of more refined component movements.

[273] In this sense, the answer advanced here could be considered a 'levels of explanation' type of response where the level of precise fine motor control is simply beyond the range of the sorts of behaviours with which we are concerned. There is a parallel between my answer to the PoP objection and Tye's characterization of the unity of bodily experience that is worth noting but that space and continuity will not allow me to elaborate upon here. See: Tye (2003, p. 47-48) for this supporting view.

constrained by AC, they nevertheless do fall under our active control.[274] It seems clear that we would at the very least be able to immediately correct any movements which might have been too loosely controlled or that may have been afforded too wide a degree of freedom with respect to the range of motion in which they were permitted by AC to operate. In short, as the quote by Davidson provided above reveals, it isn't that our coherent volitions must pick out radically precise timing and ultra-fine motor control specifications but rather, any number of loosely constrained, yet coordinated behaviours may satisfy our autonomous desires; and whatever the precise details of those actions may turn out to be, what matters to our autonomous agency as it unfolds is that we have the occurrent coherent psychological connection to and general control of our actions (i.e. in terms of what is set by our own occurrent reflective volitions) that has been proposed in section 4.4. To demand any greater precision of control from our autonomous volitions is to require something far more strict and detailed than is likely to even be consciously possible for creatures like us. In other words, the PoP not only appears to miss the target, it also sets the bar much too high.

The final complaint against our view to be considered in this section is less of an objection than it is a simple recognition that the requisite conditions for autonomy proposed by our model would appear to vastly reduce the number of cases in which people will be treated as acting autonomously in comparison to the number of cases in which most would typically consider themselves to be acting autonomously at present. The idea here is that, for instance, most people may never acknowledge as distinct from autonomy for themselves—at least not without targeted prodding and prompting—those automatic movements which by chance agree with their coherent reflective volitions. And where they may normally have believed themselves to be

[274] Contrast this point with example of the automated policy activation identified in section 1.4 and elaborated in section 4.3 above.

entirely autonomous in such cases, our model marks the failure of so simple and unconsidered an approach. Moreover, given that automaticity is a fairly common and ubiquitous phenomenon, it follows that far fewer of our movements end up being autonomous than we might generally suppose. The heart of the worry appears to be that genuinely autonomous action (according to our model), might end up being an exceedingly rare phenomenon. And relatedly, that the wide range of uses that the concept is normally put to would be radically limited or reshaped such that it could cause problems for our general understanding and attributions of other notions that are standardly assessed in terms of how they connect with autonomy; namely, notions like responsibility, consent, culpability, *et cetera*. The problem then concerns not only autonomy proper but how it relates to a number of other important practical concepts as well—concepts which are not only tightly bound, but also subject to how we understand the contours of autonomous agency.

So as to not stoke the somewhat alarmist sounding concerns raised above, I will begin by saying that the question of the frequency and extent of our automatic behaviours in the world would appear to be an empirically addressable one; and thus, I suggest that we at least provisionally exercise caution and restraint when it comes to our predictions about the extent to which treating automaticity as an impediment to autonomy might impact just how autonomous we commonly take ourselves to be.[275] It could turn out that a significant amount of our daily behaviour may be automatically performed, but it might just the same be found that automatic

[275] If we take Reason's (1979) study as a preliminary investigation into this question, we find that (as mentioned in section 2.4), although Reason relied upon an admittedly small thirty-five person sample size, participants reported on average only twelve instances of automatic behaviour over the two week trial period. That's less than one recognized instance of automaticity a day, which doesn't seem all that threatening. Of course, one important thing to point out is that Reason's study was only concerned with those automatic behaviours which led to some error of action (i.e. actions not as planned). So, given that Reason's study relied upon a small sample size, participant self-reports, and only focused upon automatic behaviours that amounted to some recognized error, one might remain reasonably unconviced that such research could be representative of normal levels of daily instances of automaticity. Nevertheless, the study may help to at least somewhat relax concerns about the scope of the impact of automatic behaviours upon autonomy.

behaviours affect only a small fraction of our daily lives and perhaps such behaviours are primarily restricted to highly routine tasks at that.[276] Given that we are still awaiting further empirical data on the reach and frequency of automaticity, I think that we ought to remain prudent against inflating within our imaginations just how radical the proposal of this dissertation may be.

The above said, even if it were to be discovered that a significant amount of our daily behaviour turned out to be accomplished automatically, I would not withdraw the proposal advanced herein. The model of autonomy that has been developed in this chapter is, I believe, more carefully consistent with and responsive to the underlying notion of active self-governance than any previous view. And if what we want in our model of autonomy is an honest account of active self-governance, then we must acknowledge the worries that led to its construction. Moreover, there are a number of other already recognized constraints upon autonomy that may likewise translate into autonomous actions being potentially very infrequent, and many of these constraints may have far more ambiguous boundaries. For instance, Dworkin's (1988) identification of coercion (mentioned in section 1.2.2) as a factor that may undermine an individual's autonomy could be characterized in terms of say, the result of a specific direct personal threat. However, it may alternatively be maintained that things like the political and economic systems within which people generally live and operate are equally coercive and constraining upon behaviour; and thus, that they equally undermine autonomy (and do so in a very broad way at that). Indeed if one accepts the latter view, one will already be inclined to think that autonomy is exceedingly rare. And if it in fact turns out that automaticity is so widespread that it forces us to re-examine other notions that are deeply connected with or

[276] Such a view would be consistent with what is known with respect to the common conditioning component of automaticity (as was covered in section 2.1).

dependent upon personal autonomy, then I think that we will simply have to reconsider our evaluations and understandings of those notions. Pretending that automaticity is not a genuine problem will not improve the accuracy of application of those types of dependant notions. Nevertheless, even if the data ends up supporting the view that automaticity is currently very widespread, there are reasons to think that things might not remain so.

One benefit of accepting the model of autonomy advanced in this chapter is that, the more widely recognized the criteria for genuinely autonomous actions become, the more people are empowered to actively counteract instances of automaticity by recognizing the value of putting in the effort to remain attentively engaged in what they are doing as well as having acted from coherent reflective volitions and with a particular form of unified consciousness. And I don't believe that such a suggestion is akin to wishful thinking. Indeed, there exist a number of what have often been referred to as "mindfulness" practices that appear to both hold a broad appeal, and are often also of an ancient pedigree. What this suggests to me is that, when people are introduced to ideas and techniques about how to become more consciously aware of themselves and their surroundings, they tend to value the outcomes of such practices enough to continue with them. And given that the model of autonomy that has been advanced in this chapter is essentially a form of active (perhaps non-meditative) "mindfulness", I don't see why it would be incapable of holding an at least equal level of general appeal and adoption once disseminated. And therefore, with this knowledge in hand, we may begin to counteract just how widespread instances of automaticity may be.

For the above mentioned reasons, I don't believe that we ought to be very concerned about what might follow from adopting the model of autonomy proposed in this dissertation (i.e. in terms of fears about a significant reduction to how autonomous we in fact are). In the next and

final section of this chapter, I will close with some final comments and suggestions for future research.

4.6 Final Comments

While this dissertation has certainly covered a lot of ground, beginning with a review and revamping of an internalist approach to autonomous agency, to elaborating a newly encountered significant threat to autonomy (i.e. automaticity), before finally considering how to address that problem, it is unlikely that the proposal advanced here will put an end to discussion and debate about the nature, structure, and impact of autonomy. And this is as it should be. Each of the major theoretical views that we have considered along the way has had something to contribute to the refinement of thinking upon the notion of autonomy, and although the contribution made by this dissertation may amount to a further degree of refinement in that thought, it is next to certain to not have had the last word on the matter. Given that the research upon and thinking about autonomy is likely to continue to develop over time, I would like to say some things about some areas of focus that I think are likely to be of significant benefit.

One of the things that I aspire to have contributed to with this dissertation is an increased recognition of the value of empirical research to theorizing about autonomy. My work here is by no means the first to have embraced this sort of interdisciplinary approach but, the hope is that it might encourage further contact between research done in psychology and that done within philosophy. I believe that there are many fertile areas of research still waiting to be explored and developed at the intersection of psychological findings and philosophical theorizing. In fact,

many of the subsections of chapters two and three dealing with specific areas of psychological research could stand to be explored much more thoroughly than could be done here; areas like the research on habit, or attention as selection for action for example. And there are a number of other underexplored but equally interesting areas of psychological research that appear to have implications for autonomy; research on well-known phenomena like cognitive dissonance for example, or obedience studies. If the work done here at all stimulates further research along such lines, I will consider it to have been worthwhile. But as philosophy itself continues to splinter into distinct areas of specialization, I think that continuing to make connections between certain related specialized areas of purely philosophical research also holds great potential. Specifically, as was done in this dissertation, I think that looking for points of contact between the sub-disciplines of research on agency or action theory and research on consciousness holds the potential for mutual impact and significant theoretical advancement. One specific area to mention here would be a more thorough consideration of how the different sorts of background states of consciousness might impact autonomous agency. As should be clear, there exists a number of exciting avenues open to further or fresh investigation before us. Looking back and recognizing just how far our thinking and research has brought us can be a source of great inspiration and encouragement to embark upon new future work. It has been a privilege to have made contact with the work that preceded this dissertation as it will be to continue to follow the future research on personal autonomy.

BIBLIOGRAPHY

Aarts, H & Dijksterhuis, A. (2000). The automatic activation of goal-directed behaviour: The case of travel habit. *Journal of Environmental Psychology, 20*, 75-82.

Allport, A. (1987). Selection for action: some behavioral and neurophysiological considerations of attention and action. In H. Heuer & a. F. Sanders (1987). (Eds.), *Perspectives on perception and action*. (pp. 395-419). New Jersey: Lawrence Erlbaum Associates.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology, 24* (2), 225-235.

Anscombe, G. E. M. (1972) *Intention*. Oxford: Basil Blackwell. 2nd edn. Reprint.

Anscombe, G. E. M. (1989). Von Wright on practical inference. In P. A. Schlipp and L. E. Hahn (Eds.), *The philosophy of George Henrik Von Wright*. (pp. 377-404). Illinois: Open Court.

Aristotle. Nicomachean ethics. In J.D. Kaplan (Ed.), (1958). *The pocket Aristotle* (pp. 158-274). New York: Simon & Schuster.

Armstrong, K. M., Schafer, R. J., Chang, M. H. & Moore, T. (2012). Attention and action in the frontal eye field. In G. R. Mangun (2012). (Ed.), *The neuroscience of attention: attention control and selection*. (pp. 151-166). New York: Oxford University Press.

Arpaly, N. (2004). Which Autonomy? In J. K. Campbell, M. O'Rourke & D. Shier (Eds.), *Freedom and Determinism* (pp. 173–188). Cambridge: MIT Press.

Attneave, F. (1960). In defence of humunculi. In W. Rosenblith (Ed.), *Sensory communication* (pp. 777–782). Cambridge: MIT Press.

Awh, E. & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences, 5* (3), 119-126.

Baars, B. J. (1997). Some essential differences between consciousness and attention, perception, and working memory. *Consciousness and Cognition, 6*, 363-371.

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences, 6* (1), 47-52.

Baddeley, A. D. (1981). The concept of working memory: A view of its current state and probable future development. *Cognition, 10*, 17-23.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D. (1990). *Human memory: theory and practice*. Massachusetts: Allyn and Bacon.

Baddeley, A. D. (1992). Working memory. *Science, 255,* 556-559.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 49A* (1), 5-28.

Baddeley, A. D. (1998). The central executive: A concept and some misconceptions. *Journal of the International Neuropsychological Society, 4*, 523-526.

Baddeley, A. D. (2002a). Is working memory still working?. *European Psychologist, 7* (2), 85-97.

Baddeley, A. D. (2002b). Fractionating the central executive. In D. T. Stuss & R. T. Knight, (Eds.), *Principles of frontal lobe function*. (pp. 246-260). New York: Oxford University Press.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews: Neuroscience, 4*, 829-839.

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G.A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–90). New York: Academic Press.

Baeroe, K. (2010). Patient autonomy, assessment of competence and surrogate decision-making: A call for reasonableness in deciding for others. *Bioethics, 24* (2), 87-95.

Bandura, A. (1977a). Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review, 84,* 191-215.

Bandura, A. (1977b). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1990). Self-regulation of motivation through anticipatory and self-reactive mechanisms. In R. A. Dienstbier (Ed.), *Perspectives on motivation: Nebraska symposium on motivation* (Vol. 38, pp. 69-164). Lincoln: University of Nebraska Press.

Bandura, A. (1997). *Self-efficacy*. New York: Freeman.

Bargh, J. A. (1989). Conditional automaticity : Varieties of automatic influence in social perception and cognition. In S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3-51). New York: Guilford Press.

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of social interaction. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition* (Vol. 2, pp. 93-130). New York: Guilford Press.

Bargh, J. A. (1992). The ecology of automaticity: toward establishing the conditions needed to produce automatic processing effects. *The American Journal of Psychology, 105* (2), 181-199.

Bargh, J. A. (1994). The Four Horsemen of automaticity : awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1-40). Hillsdale, NJ: Erlbaum.

Bargh, J. A. (1996). Principles of automaticity. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: handbook of basic principles* (pp. 169-183). New York: Guilford.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *The automaticity of everyday life: advances in social cognition* (Vol. 10, pp. 1-61). Mahwah, NJ: Erlbaum.

Bargh, J. A. (2005). Bypassing the will: toward demystifying the nonconscious control of social behaviour. In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.), *The New Unconscious*. New York: Oxford University Press.

Bargh, J. A. & Barndollar, K. (1996). Automaticity in action: the unconscious as repository of chronic goals and motives. In P.M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: linking cognition and motivation to behaviour.* (pp. 457-481). New York: Guilford Press.

Bargh, J. A. & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54* (7), 462-479.

Bayne, T. (2005). Divided brains and unified phenomenology: A review essay on Michael Tye's consciousness and persons. *Philosophical Psychology, 18* (4), 495-512.

Bayne, T. (2010). *The unity of consciousness*. Oxford: Oxford University Press.

Bayne, T. & Chalmers, D. J. (2003). What is the unity of consciousness? In A. Cleeremans (Ed.), *The Unity of Consciousness: Binding, Integration, Dissociation*. (pp. 23-58). New York: Oxford University Press.

Berlyne, D. E. (1960). *Conflict, arousal and curiosity*. New York: McGraw-Hill.

Berofsky, B. (1995). *Liberation from Self: A Theory of Personal Autonomy*, New York: Routledge and Kegan Paul.

Block, N (1995). On a confusion about a function of consciousness. *Behavioral and Brain Science*s, 18, 227-247.

Block, N (1997). On a confusion about a function of consciousness. In N. Block, O. Flanagan & G. Guzeldere (Eds.), *The Nature of Consciousness: Philosophical Debates*. (pp. 375-415). Cambridge: MIT Press.

Bratman, M. E. (1981). Intention and means-end reasoning. *Philosophical Review, 90* (2), 252-265.

Bratman, M. E. (1989). Intention and personal policies. *Philosophical Perspectives, 3*, 443-469.

Bratman, M. E. (1991). Cognitivism about practical reason. *Ethics, 102* (1), 117-128.

Bratman, M. E. (1992). Practical reasoning and acceptance in a context. *Mind, 101* (401), 1-15.

Bratman, M. E. (2000). Reflection, planning, and temporally extended agency. In M. E. Bratman (2007). (Ed.), *Structures of Agency*. (pp. 21-46). New York: Oxford University Press.

Bratman, M. E. (2002). Hierarchy, circularity, and double reduction. In M. E. Bratman (2007). (Ed.), *Structures of Agency*. (pp. 68-88). New York: Oxford University Press.

Bratman, M. E. (2004). Three theories of self-governance. In M. E. Bratman (2007). (Ed.), *Structures of Agency*. (pp. 222-253). New York: Oxford University Press.

Bratman, M. E. (2005). Planning agency, autonomous agency. In M. E. Bratman (2007). (Ed.), *Structures of Agency*. (pp. 195-221). New York: Oxford University Press.

Bratman, M. E. (2007). Anchors for deliberation. In C. Lumer, S. Nannini (2007). (Eds.), *Intentionality, Deliberation and Autonomy: The Action-Theoretic Basis of Practical Philosophy*. (pp. 187-205). Vermont: Ashgate Publishing Company.

Braun, J. (1998). Divided attention: narrowing the gap between brain and behavior. In R. Parasuraman (1998). (Ed.), *The Attentive Brain*. (pp. 327-351). Cambridge: MIT Press.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. Journal of Experimental Psychology, 47 (3), 191-196.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review, 64* (3), 205-215.

Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon Press.

Broadbent, D. E. (1982). Task combination and selective intake of information. *Acta Psychologica, 50*, 253-290.

Brody, H. (1985). Autonomy revisited: Progress in medical ethics: Discussion paper. *Journal of the Royal Society of medicine, 78,* 380-387.

Brook, A. & Raymont, P. (2006). The unity of consciousness. Stanford Encyclopedia of Philosophy. Retrieved Aug. 3rd 2008 from: http://plato.stanford.edu/entries/consciousness-unity/

Brudner, A. (2000). Insane automatism: a proposal for reform. *McGill Law Journal, 45* (1), 65-86.

Buss, S. (2002). Personal autonomy. Stanford Encyclopedia of Philosophy. Retrieved Aug 3rd 2008 from: http://plato.stanford.edu/entries/personal-autonomy/

Chanon, V. W. & Hopfinger, J. B. (2008). Memory's grip on attention: the influence of item memory on the allocation of attention. *Visual Cognition, 16* (2/3), 325-340.

Chen, S., Fitzsimons, G. M. & Andersen, S. M. (2007). Automaticity in close relationships. In J. A. Bargh (2007). (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes*. (pp.133-172). New York: Psychology Press.

Christman, J. (1987). Autonomy: A defense of the split-level self. *Southern Journal of Philosophy, 25* (3), 281-293.

Christman, J. (1988). Constructing the inner citadel: Recent work on the concept of autonomy. *Ethics, 99* (1), 109-124.

Christman, J. (1989). *The inner citadel: Essays on individual autonomy*. New York: Oxford University Press.

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy, 21* (1), 1-24.

Christman, J. (1993). Defending historical autonomy: A reply to professor Mele. *Canadian Journal of Philosophy, 23* (2), 281-289.

Clark, F., Sanders, K., Carlson, M., Blanche, E. & Jackson, J (2007). Synthesis of habit theory. *OTJR: Occupation, Participation, and Health,* 27(Suppl), 7S-23S.

Cohen, R. A., Sparling-Cohen, Y. A. & O'Donnell, B. F. (1993). *The neuropsychology of attention*. New York: Plenum Press.

Cooper, R. J. (1994). Automatism. *Journal of Criminal Law, 58* (2), 162-163.

Cuypers, S. E. (2000). Autonomy beyond voluntarism: In defense of hierarchy. *Canadian Journal of Philosophy, 30* (2), 225-256.

Cuypers, S. E. & Haji, I. (2006). Education for critical thinking: Can it be non-indoctrinative? *Educational Philosophy and Theory, 38* (6), 723-743.

Dainton, B. (2006). *Stream of Consciousness*. New York: Routledge.

Dashiell, J. F. (1928). *Fundamentals of objective psychology*. New York: Houghton Mifflin.

Davidson, D. (1963). Actions, reasons, and causes. In D. Davidson (1980). (Ed.), *Essays on actions and events*. (pp. 3-19). New York: Oxford University Press.

Davidson, D. (1973). Freedom to act. In D. Davidson (1980). (Ed.), *Essays on actions and events*. (pp. 63-81). New York: Oxford University Press.

Davidson, D. (1978). Intending. In D. Davidson (1980). (Ed.), *Essays on actions and events*. (pp. 83-102). New York: Oxford University Press.

Dearden, R. F. (1972). Autonomy and Education. In R. F. Dearden, P. H. Hirst & R. S. Peters (Eds.), *Education and Reason. Part 3 of Education and the Development of Reason.* (pp. 58-75). London: Routledge & Kegan Paul.

De Biran, M (1929). *The influence of habit on the faculty of thinking*. Baltimore: Waverly Press.

Deutsch, J. A. & Deutsch, D. (1963). Attention: some theoretical considerations. *Psychological Review, 70* (1), 80-90.

Dijksterhuis, A., Chartrand, T. L., & Aarts, H. (2007). Effects of priming and perception on social behaviour and goal persuit. In J. A. Bargh (Ed.), Social psychology and the unconscious: the automaticity of higher mental processes. (pp. 51-131). New York: Psychology Press.

Dimock, S. (1997). Personal autonomy, freedom of action, and coercion. In S. Brennan, T. Issacs & M. Milde (Eds.), *A Question of Values: New Canadian Perspectives on Ethics and Political Philosophy.* (pp. 65-86). Amsterdam: Rodopi Press.

Dimock, S. (2011). What are intoxicated offenders responsible for? The "intoxication defense" re-examined. *Criminal Law and Philosophy, 5* (1), 1-20.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology, 92* (1), 53-78.

Dworkin, G. (1970). Acting freely. *Nous, 4* (4), 367-383.

Dworkin, G. (1976). Autonomy and behaviour control. *The Hastings Center Report, 6* (1), 23-28.

Dworkin, G. (1981). The concept of autonomy. In J. Christman (1989). (Ed.), *The Inner Citadel:Essays on individual autonomy*. (pp. 54-62). New York: Oxford University Press.

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Ericsson, K. A., Krampe, R. T. & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Feinberg, J. (1986). *The moral limits of the criminal law Vol. 3: Harm to self*. New York: Oxford University Press.

Festinger, L. (1957). *A theory of cognitive dissonance.* California: Stanford University Press.

Festinger, L. & Carlsmith, J. M. (1959) Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203-211.

Fischer, J.M. & Ravizza, M. (1993). *Perspectives on Moral Responsibility*. Ithaca: Cornell University Press.

Fougnie, D. & Marois, R. (2006). Distinct capacity limits for attention and working memory: evidence from attentive tracking and visual working memory paradigms. *Psychological Science, 17* (6), 526-534.

Franconeri, S. L., Alvarez, G. A. & Cavanagh, P. (2013). Flexible cognitive resources: competitive content maps for attention and memory. *Trends in Cognitive Sciences, 17* (3), 134-141.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. In H. Frankfurt (1988). (Ed.), *The importance of what we care about*. (pp. 11-25). New York: Cambridge University Press.

Frankfurt, H. (1982). The importance of what we care about. In H. Frankfurt (1988). (Ed.), *The Importance of What We Care About*. (pp. 80-94). New York: Cambridge University Press.

Frankfurt, H. (1984). Necessity and desire. In H. Frankfurt (1988). (Ed.), *The Importance of What We Care About*. (pp. 104-116). New York: Cambridge University Press.

Frankfurt, H. (1987). Identification and wholeheartedness. In H. Frankfurt (1988). (Ed.), *The Importance of What We Care About*. (pp. 159-176). New York: Cambridge University Press.

Frankfurt, H. (1992). The faintest passion. *Proceedings and Addresses of the American Philosophical Association, 66* (3), 5-16.

Frankfurt, H. (1993). Autonomy, necessity, and love. In H. Frankfurt (1999). (Ed.), *Necessity, volition, and love*. (pp. 129-141). New York: Cambridge University Press.

Friedman, M. (1986). Autonomy and the split-level self. *Southern Journal of Philosophy, 24* (1), 19-35.

Glaser, J. & Kihlstrom, J. F. (2005). Compensatory automaticity: unconscious volition is not an oximoron. In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.), *The New Unconscious*. New York: Oxford University Press.

Hancock, P. A., Oron-Gilad, T. & Szalma, J. L. (2007). Elaborations of the multiple-resource theory of attention. In A. F. Kramer, D. A. Wiegman & A. Kirlik (2007). (Eds.), *Attention: from theory to practice*. (pp. 45-56). New York: Oxford University Press.

Harmon-Jones, E. & Mills, J. (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington: American Psychological Association

Healy, P. (2000). Automatism confined. *McGill Law Journal, 45* (1), 87-106.

Helmholtz, H. von (1867/1925). *Treatise on physiological optics* (translated from the 3rd German edition). New York: Dover.

Hirst, W., Spelke, E. S., Reaves, C. C., Caharack, G. & Neisser, U. (1980). Dividing attention without alternation or automaticity. *Journal of Experimental Psychology, 109*, (4), 98-117.

Holland, W. H. (1982). Automatism and criminal responsibility. *Criminal Law Quarterly, 25* (1), 95-128.

Hommel, B. (2010). Grounding attention in action control: the intentional control of selection. In B. Bruya. (2010). (Ed.), *Effortless attention*. (pp. 212-40). Cambridge: MIT Press.

Hurley, S. (1994). Unity and objectivity. In C. Peacocke (Ed.), *Objectivity, Simulation and the Unity of Consciousness: Current Issues in the Philosophy of Mind*. (pp. 49-77) Oxford: Oxford University Press.

Hurley, S. (1998). *Consciousness in Action*. Massachusetts: Harvard University Press.

James, W. (1890). *The principles of psychology* (2 Vols.). New York: Holt.

Jastrow, J. (1906). *The subconscious*. Boston, MA: Houghton-Mifflin.

Kahneman, D. (1973). *Attention and effort*. New Jersey: Prentice-Hall.

Kalant, H. (1996). Intoxicated automatism: legal concept vs. scientific evidence. *Contemporary Drug Problems, 23* (4), 631-648.

Keele, S. W. (1973). *Attention and human performance*. California: Goodyear.

Kimberg, D. Y., D'Esposito, M. and Farah, M. J. (1997). Cognitive functions in the prefrontal cortex—Working memory and executive control. *Current Directions in Psychological Science, 6* (6), 185-192.

Klumb, P. L. (1995). *Attention, action, absent-minded aberrations*. Berlin: Peter Lang.

Koch, C. & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Science, 11* (1), 16-22.

Kukla, R. (2005). Conscientious autonomy: Displacing decisions in health care. *The Hastings Center Report, 35* (2), 34-44.

Kunda, Z. (1999). *Social cognition: Making sense of people.* Cambridge: MIT Press.

Mark, M. M., Sinclair, R. C., & Wellens, T. R. (1991). The effect of completing the Beck Depression Inventory on self-reported mood state: Contrast and assimilation. *Personality and Social Psychology Bulletin, 17*, 457-465.

McCann, H. J. (1998). *The works of agency: On human action, Will, and freedom*. Ithaca: Cornell University Press.

McLeod, P., McLaughlin, C. & Nimmo-Smith, I. (1985). Information encapsulation and automaticity: evidence from the visual control of finely timed actions. In M. I. Posner & O. S. M. Marin (1985). (Eds.), *Attention and performance XI*. (pp. 391-406). New Jersey: Lawrence Erlbaum Associates.

McSherry, B. (2005). Men behaving badly: Current issues in provocation, automatism, mental impairment and criminal responsibility. *Psychiatry, Psychology and Law, 12* (10), 15-21.

Mele, A. (1993). History and personal autonomy. *Canadian Journal of Philosophy, 23* (2), 271-280.

Mele, A. (2003). *Motivation and Agency*. New York: Oxford University Press.

Miller, B. L. (1981). Autonomy & the refusal of lifesaving treatment. *The Hastings Center Report, 11* (4), 22-28.

Mole, C. (2008). Attention and Consciousness. *Journal of Consciousness Studies, 15* (4), 86-104.

Mole, C. (2009). Attention. Stanford Encyclopedia of Philosophy. Retrieved Sept 4th 2012 from: http://plato.stanford.edu/entries/attention/

Mole, C. (2011). *Attention is cognitive unity: an essay in philosophical psychology*. New York: Oxford University Press.

Moray, N. (2007). Attention from history to application. In A. F. Kramer, D. A. Wiegman & A. Kirlik (2007). (Eds.), *Attention: from theory to practice*. (pp. 3-15). New York: Oxford University Press.

Nagel, T. (2003). Freedom. In G. Watson, (Ed.). *Free Will*. (pp. 229-256). Oxford: Oxford University Press.

Navon, D. (1985). Attention division or attention sharing? In M. I. Posner & O. S. M. Marin (1985). (Eds.), *Attention and performance XI*. (pp. 133-146). New Jersey: Lawrence Erlbaum Associates.

Neumann, O. (1987). Beyond capacity: a functional view of attention. In H. Heuer & a. F. Sanders (1987). (Eds.), *Perspectives on perception and action*. (pp. 361-394). New Jersey: Lawrence Erlbaum Associates.

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75* (6), 522-536.

Norman, D. A. & Shallice, T. (1980). Attention to action: Willed and automatic control of behaviour. *CHIP Report 99*. San Diego: University of California.

Norman, D. A. & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz & D. Shapiro (1986). (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 3-18). New York: Plenum Press.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A. & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience, 5* (11), 1203-1209.

Oh, S-H. & Kim, M-S. (2004). The role of spatial working memory in visual search efficiency. *Psychonomic Bulletin & Review, 11* (2), 275-281.

Olivers, C. N. L. (2010). The attentional boost and the attentional blink. In A. C. Nobre & J. T. Coull (2010). (Eds.), *Attention and time*. (pp. 49-62). New York: Oxford University Press.

Parasuraman, R. & Davies, D. R. (Eds). (1984). *Varieties of attention*. Orlando: Academic Press.

Pashler, H. E. (1998). *The psychology of attention***.** Cambridge: MIT Press.

Pilsbury, W. B. (1973). *Attention*. New York: Arno Press.

Posner, M. I. (1994). Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Science, 91*, 7398-7403.

Posner, M. I. (2008). Attentional networks and the semantics of consciousness. *Psyche: An Interdisciplinary Journal of Research on Consciousness,14*, 1-7.

Posner, M. I. (2012). *Attention in a social world*. New York: Oxford University Press.

Prinz, J. (2000). The ins and outs of consciousness. *Brain and Mind, 1*, 245-256.

Reason, J. (1979). Actions not as planned: The price of automatization. In G. Underwood (1993). (Ed.), *The psychology of attention* (Vol. 2, pp. 150-172). New York: New York University Press.

Richards, D. A. J. (1989). Autonomy in law. In J. Christman (1989). (Ed.), *The inner citadel: Essays on individual autonomy*. (pp. 246-258). New York: Oxford University Press.

Robertson, I. H. & O'Connell, R. (2010). Vigilant attention. In A. C. Nobre & J. T. Coull (2010). (Eds.), *Attention and time*. (pp. 79-88). New York: Oxford University Press.

Schall, J. D. & Woodman, G. F. (2012). A stage theory of attention and action. In G. R. Mangun (2012). (Ed.), *The neuroscience of attention: attention control and selection*. (pp. 187-208). New York: Oxford University Press.

Sellars, M. (2007). *Autonomy in the law*. Dordrecht: Springer.

Shallice, T. (2002). Fractionation of the supervisory system. In D. T. Stuss & R. T. Knight, (Eds.), *Principles of frontal lobe function*. (pp. 261-277). New York: Oxford University Press.

Shapiro, K. L. & Raymond, J. E. (2010). The attentional blink: temporal constraints on consciousness. In A. C. Nobre & J. T. Coull (2010). (Eds.), *Attention and time*. (pp. 35-48). New York: Oxford University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Siegel, H. (1997). *Rationality redeemed?: Further dialogues on an educational ideal.* New York: Routledge.

Skelton, J. A., & Strohmetz, D. B. (1990). Priming symptom reports with health related cognitive activity. *Personality and Social Psychology Bulletin, 16*, 449-464.

Strawson, G. (1997). The self. In R. Martin & J. Barresi (2003). (Eds.), *Personal Identity*. (pp. 335-377). Massachusetts: Blackwell Publishing.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18* (6), 643-662.

Taylor, J. S. (2005). *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*. Cambridge: Cambridge University Press.

Thalberg, I. (1978). Hierarchical analyses of unfree action. *Canadian Journal of Philosophy, 8* (2), 211-226.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12* (4), 242-248.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological review, 76* (3), 282-299.

Tye, M. (2003). *Consciousness and persons: unity and identity*. Massachusetts: MIT Press.

Underwood, G (Ed.). (1993). *The psychology of attention vol.1&2*. New York: New York University Press.

Velleman, J. D. (1992). What happens when someone acts?. *Mind, 101* (403), 461-481.

Velleman, J. D. (1996). The possibility of practical reason. *Ethics, 106* (4), 694-726.

Velleman, J. D. (2000a). *The possibility of practical reason*. New York: Oxford University Press.

Velleman. J. D. (2000b). From self-psychology to moral philosophy. In J. E. Tomberlin (Ed.), *Philosophical perspectives: #14 Action and freedom*. (pp. 349-377). Massachusetts: Blackwell Publishers.

Velleman, J. D. (uploaded 2007). The way of the wanton. Retrieved May 10th 2008 from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1006893

Watson, G. (1975). Free agency. In J. Christman (1989). (Ed.), *The inner citadel: Essays on individual autonomy*. (pp. 109-122). New York: Oxford University Press.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge: MIT Press.

White, J. (1991). *Education and the good life: Autonomy, altruism, and the national curriculum.* New York: Teachers College Press.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (1984). (Eds.), *Varieties of attention*. (pp. 63-102). Florida: Academic Press.

Wickens, C. D. (2007). Attention to attention and its applications: a concluding view. In A. F. Kramer, D. A. Wiegman & A. Kirlik (2007). (Eds.), *Attention: from theory to practice*. (pp. 239-249). New York: Oxford University Press.

Winch, C. (2006). *Education, autonomy and critical thinking.* New York: Routledge.

Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.

Wolters, G. & Prinsen, A. (1997). Full versus divided attention and implicit memory performance. *Memory & Cognition, 25* (6), 764-771.

Wood, W. & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review, 114* (4), 843-863.

Woodman, G. F. & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic Bulletin & Review, 11* (2), 269-274.

Wright, R. D. & Ward, L. M. (2008). *Orienting of attention*. New York: Oxford University Press.

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein & T. Vierkant (2013). (Eds.), *Decomposing the will*. (pp. 244-261). New York: Oxford University Press.