Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Julia Haas Date

Weakness of Will: A Case Study for Moral Philosophy and the Cognitive Neurosciences

By

Julia Haas

Doctor of Philosophy

Philosophy

_____

Richard Patterson
Advisor

_____

Robert N. McCauley
Committee Member

_____

Ursula Goldenbaum
Committee Member

_____

Mark Risjord
Dissertation Reader

_____

Gregory S. Berns
Dissertation Reader

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

Weakness of Will: A Case Study for Moral Philosophy and the Cognitive Neurosciences

By

Julia Haas

B.A., Concordia University, 2008
M.A., Emory University, 2011

Advisor: Richard Patterson, PhD.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Philosophy
2014

Abstract

Weakness of Will: A Case Study for Moral Philosophy and the Cognitive Neurosciences
By Julia Haas

My doctoral dissertation, "Weakness of Will: A Case Study for Moral Philosophy and the Cognitive Neurosciences," provides a naturalistic theory of why agents act against their better judgment. Drawing on evidence from computational modeling and cognitive neuroscience, I demonstrate that suboptimal interactions between three known decision-making controllers (i.e., the Pavlovian, goal-directed, and habit-based controllers) elicit different, albeit psychologically indistinguishable kinds of weakness of will.

Weakness of Will: A Case Study for Moral Philosophy and the Cognitive Neurosciences

By

Julia Haas

B.A., Concordia University, 2008
M.A., Emory University, 2011

Advisor: Richard Patterson, PhD.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Philosophy
2014

Acknowledgements

I would first and foremost like to thank Richard Patterson for his invaluable advice and support throughout the preparation of this dissertation. Without fail, his questions and suggestions challenged me to become a better and more careful philosopher. He also patiently read and re-read every last section of the dissertation. I could not have asked for a better supervisor. I would also like to thank Cynthia Patterson for her generosity and support. Robert McCauley first introduced me to naturalism and I never looked back. He has been a wonderful teacher and advisor and in many respects set me on my path. Ursula Goldenbaum first taught me about philosophy, science, and, of course, Spinoza. She has been a wonderful mentor throughout graduate school. I am grateful to my readers Mark Risjord and Gregory Berns for their insightful suggestions and criticisms. Ann Hartle always encouraged me and taught me a great deal. Catherine Hall, Michael Hodgin, Debbie Miller, and Frances Campbell made everything work for six straight years. The Laney Graduate School provided me with all of the support I could have asked for.

Karen Rommelfanger and Gillian Hue were generous and encouraging mentors. Patricia Churchland and William Casebeer treated me like a sensible young academic at just the right moment. I would like to thank Peter Dayan for inviting me to speak at the Gatsby Computational Neuroscience Unit and generously sitting down with me to talk about my ideas. I had wonderful conversations with Molly Crockett, Andreas Hula and Xiaosi Gu.

Carolina Campanella, Constance Scott Harrell Schreckengost, Jill Carpenter-Smith, and Brian Dias patiently taught me about science, statistics, and running. I met Edward Glowienka III, Catherine Homan, Timothy Harfield, Rebecca Longtin Hansen, and Samuel Timme in my first year of graduate school, and they were and are my friends throughout. Kristina Gupta was my academic role model for a long time and became a dear friend (and a mom!) in the process.

I would like to thank my family for their help and encouragement: George Haas, Zora Zeman, Šárka Friedl, Barbara Haas, and Martin Friedl. Adela Maciejewski Scheer's faith in me never wavered. She always said, 'You can do it!' Often, saying it makes it so.

This dissertation could not have been written without my partner, Craig Henchey. Throughout, he never once failed to hear out my latest theory or idea. I am deeply grateful for his continuing love, wisdom, and support. I look forward to many more years of discussing philosophy with him.

I dedicate this dissertation to Vladimir Zeman. His life was a testament to the phrase, 'Be a philosopher; but, amidst all your philosophy, be still a man.'

Table Of Contents

Chapter 5

Behavioral And Neuroscientific Evidence Supporting The Habit-Based And Pavlovian Cognitive And Behavioral Hypotheses Of Weakness Of Will

Chapter 6

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Boxes

Chapter 4

PART I

HISTORICAL PERSPECTIVES ON WEAKNESS OF WILL

My box of tools was different from everyone else's,
and they had tried all their tools on it before giving the problem to me.
- R. Feynman, *Surely You're Joking, Mr. Feynman*

CHAPTER 1

SEARCHING FOR THE MECHANISMS UNDERLYING WEAKNESS OF WILL

1. Introduction: Mechanisms and Breakdowns

A mechanism is a system of interacting parts and processes. It produces a specific effect or set of effects. Distinct starting conditions and consistent processes bring about the uniform finishing conditions (Machamer, Darden, and Craver 2000; Darden 2002). For example, a ball and socket joint is a kind of mechanism, where the ball- and cup-shaped parts rotate to produce motion along a range of axes. In the case of a shoulder (Figure 1.1.), the starting conditions correspond to the neutral position of the joint. The processes correspond to the flexing of muscles and the rotation of the scapula (Brinckmann *et al.* 2002, 134). The finishing conditions correspond to the upper arm arriving in its new position.

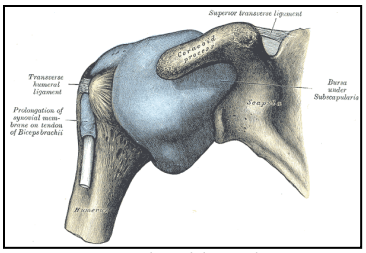

Figure 1.1. Shoulder joint (Gray's Plate 327)

To search for a mechanism is to try and understand how a given set of entities and processes works. Although discoveries can sometimes be represented as 'aha' moments, as in the famous case of Archimedes at the baths, or even come on the heels of a fortuitous mistake, as with the invention of Saran (later sold as Saran Wrap) (Yam 2009, 100), most theoretical discoveries come about as the products of systematic strategies involving hypotheses, testing, and revision (Craver and Darden, 2013, 64).

Many philosophical accounts of scientific explanation, influenced by the deductive-nomological account of explanation, emphasize the importance of theories and laws. These kinds of explanations emphasize abstract principles and general applicability. In the philosophy of action, for example, it is common to discuss explanations in terms of laws

under which observed behaviors can be subsumed (Bechtel and Abrahamsen 2005). By contrast, in the search for mechanisms, the scope of inquiry is typically narrower, with researchers focusing "from the beginning on the specifics of the composition and organization of a mechanism that generates a particular form of behavior" (Bechtel 2008, 4). The process often gradually moves over time from a lack of understanding to a clear grasp of the mechanism. Mechanisms can described using visual, linguistic, mathematical, and/or graphical representations (Craver and Darden, 2013, 38-50). The representations can also be referred to as 'models' or 'schemas.'[1]

Nevertheless, as philosopher of science William Bechtel observes, the search for mechanisms can be difficult and time-consuming. He writes, "discovering such functional components in natural biological systems is never easy. In normally operating systems the identity of the components is concealed by their smooth coordination in performing the system's overall activities. Often this smooth coordination involves non-linear interactions of large numbers of components" (2002, 229). In an effort to penetrate this 'smooth coordination,' one common catalyst in the search for mechanisms is trying to understand what happens when a system breaks down.[2] To stay with the example of the shoulder, for example, a child may take a doll whose arm has fallen off and try to figure how the limbs fit back into the socket. Along very similar lines, Hippocrates used the observation that shoulders only dislocate in one direction, namely, downward (never upward or outward), to

[1] Craver and Darden distinguish between mechanism 'sketches' and 'schemas,' where the former represent preliminary, incomplete estimates of how a system might work, while the latter correspond to increasingly comprehensive and accurate accounts of the mechanism in question (2013, 31).

[2] There are of course other ways to try to penetrate the 'smooth coordination' of complex mechanisms. R. Patterson notes that in many cases, it is more efficient to take a system apart and (with luck) to put it back together again than it is to wait for it to break down. This happens when, e.g., a novice engineer takes apart a radio or iron to try and see how it works and then reassembles it again. Frequently, the task of reassembling the mechanism is at least as informative as taking it apart is. Other methods of discovery include considering other, better understood mechanisms, as well as what Craver and Darden call 'forward/backward chaining,' that is, trying to figure out what will happen in the early and late stages of a mechanism's deployment (2013, 77).

hypothesize that the deltoid, the pectoral muscle, and the humerus must be attached in a particular way (*Joints*, Part 1, Page 1). Historically, malfunctioning systems of this sort have often set into motion the analysis of mechanisms, including the study of brain lesions to localize mental functions (Code 1996, 2003) and the investigation of the malignant progression of normal cells into harmful cancer cells (Weinstein and Case 2008). Part of the motivation for understanding a broken mechanism lies in the ability to repair it, but frequently, this is also an inherent interest in understanding how the part and processes work together.

Notably, the 'system breakdown' research strategy is analogous to but different from the role of anomalous phenomena that cannot be explained by a given scientific paradigm (Kuhn 1962/2012). In the former case, anomalous findings reveal inconsistencies 'one level up,' at the level of the scientific theories intended to explain the phenomena. In the latter instances, the regular workings of a mechanism break down and shed light on how its parts and processes are normally integrated. The 'breakdown' method of inquiry is especially useful in studying highly integrated, complex mechanisms, where an opaque system is temporarily 'broken open' to reveal its inner workings.

Weakness of will, or the phenomenon of acting against one's better judgment, has traditionally been investigated in the context of this 'system breakdown' research strategy. The *main* system of interest is actually the complex relationship between knowledge, motivation and action. Generally known as practical reasoning (in philosophy) or decision-making (in the natural and social sciences), this system is usually seamless and effective: on balance, we are able to evaluate the pros and cons of the alternatives in front of us, and once we make up our minds about something, we follow through on it. But practical reasoning (or decision-making) is difficult to analyze because it is underwritten by an extremely complex

network of biological functions. By contrast, an associated psychological and behavioral glitch known as weakness of will 'akrasia[3],' in which an agent does something she knows she shouldn't do, appears to be within explanatory reach. Though it is a less essential feature of our everyday lives, it provides an entry point for understanding the nature of decision-making. As a result, many of those historically interested in practical reasoning and decision-making have turned their attention to the problem of weakness of will.

We all experience moments of weakness of will. At the benign end of things, we reach for that second or third brownie when we really know we shouldn't. In more pernicious instances, we often fail to help or even harm people all the while knowing that this is not the right thing to do. More formally, 'weakness of will' can be defined as the phenomenon of acting against one's better judgment. For example, 'Barbara knows that *x* is the better thing to do, but she still does *y* instead,' or again, 'Barbara knows that eating a salad would be the better thing to do, but she still decides to eat cake instead.'

Many philosophers have taken up the challenge of accounting for the nature and underlying causes of weakness of will. Famously, Socrates offered the first philosophical examination of weakness of will in Plato's *Protagoras* and analyzed whether an individual can knowingly pursue what she knows to be a worse course of action. In what became known as one of the 'Socratic paradoxes' (Santas, 1964), Socrates argued that "no one goes willingly

[3] 'Akrasia' an 'weakness of will' will be used interchangeably throughout this dissertation. Preference will be given to 'weakness of will.' Most discussions in ancient philosophy and some discussions in modern philosophy use the term 'akrasia,' literally meaning 'lack of command.' Many medieval treatments of the issue used the term 'incontinence.' The literature since Davidson's 'How is Weakness of the Will Possible?' (1970) has used this term almost exclusively, although Holton (1999) contends at length that 'akrasia' and 'weakness of will' do not refer to the same phenomenon. I disagree, for reasons discussed in Section 2 below. I use 'weakness of will' throughout because 'akrasia' can be off-putting to readers outside of the discipline of philosophy. 'Weakness of will' is also more or less ubiquitous in the contemporary literature on weakness of will. Both of these reasons outweigh what I believe to be the term's seriously misleading connotations - mostly notably the suggestion that such a thing as 'the will' even exists.

toward the bad" (Plato, 1997, 784). Philosophers have avidly taken up the Socratic paradox and continue to discuss the issue in general into the present day.

Philosophical treatments of weakness of will typically focus on two main issues: whether instances of weakness of will genuinely exist (i.e. is it actually possible to choose what one knows to be the worse option), and, if they do, what mechanism(s) underlie them?[4] Unfortunately, and perhaps as a result of its long interpretive tradition, the debate surrounding weakness of will has become jumbled and difficult to follow. Although dozens of articles and books continue to be published on the subject each year, philosophers seem to exchange largely semantic criticisms and can come to little agreement on the subject, even in terms of just how to define the phenomenon. Many philosophical exchanges devise extensive thought experiments with the sole purpose of refuting rival definitions of weakness of will. The confusion is such that it has led a fellow philosopher, Stephen Schiffer, to observe that weakness of will does not exist as a stable concept in the literature, but is rather "an unfortunate if picturesque term of art [that] has never had better than a vacillating reference" (1976, 201). Even a cursory look at the literature makes one sympathetic to Schiffer's characterization.

2. Explaining the Phenomenon

The scholarly debate surrounding weakness of will can be made more structured and accessible, however. A preliminary degree of organization can be brought to the debate by distinguishing between the behavior of weakness of will and the many philosophical descriptions that have been offered for it. Everyday experiences of weakness of will are

[4] Levy (2012, 2) divides the latter category as the 'What' and the 'How' questions of weakness of will. He explains, "The 'what' question concerns what psychological or mental states or entities must be postulated in order to explain weakness of the will […] The 'how' question asks how the mental states or entities required to explain weakness of will actually cause the weak-willed behavior."

noncontroversial and widely accepted. It is competing descriptions and explanations of the phenomenon that have proved to be so challenging for philosophers.

The question of whether instances of it genuinely exist can, I suggest, be disposed of relatively quickly. This is because the everyday referent of the term 'weakness of will' is generally clear. We all know what it is like to do something we know we shouldn't do or hadn't wanted to do, whether it is to skip a workout or follow through on a more detrimental course of action. In many respects, even Plato's Socrates assumed that the phenomenon of weakness of will (or akrasia) was familiar enough that he could use it as an example to illustrate a more complex claim about the nature of virtue (see Chapter 3, Section 2). When he introduces the issue in the *Protagoras*, he invites his listeners to consider the familiar "experience which they call being overcome by pleasure" (Cooper, 1997/353a). Problems only arise once the group begins to discuss *explanations* of the phenomenon.

Christopher Shields does a good job of characterizing this everyday behavior while carefully distinguishing it from whatever account we give of it (2007). Any discussion of weakness of will should begin, he claims, with the following "apparently incontestable datum: we sometimes resolve to pursue a course of action *a* in preference to *b*, because we suppose, or suppose that we suppose, that *a* is all-things-considered preferable to *b*, and yet then at the moment of action opt for *b*" (2007, 65). Optionally, this account can also include post-action experiences including self-recrimination, regret, and some form of resolve not to repeat this behavior in the future. Helpfully, Shields calls this experience 'implementation failure"; by using a term that is free of any historical connotations, it is more clearly and memorably distinguished from explanations of weakness of will. I follow him in this usage, and we can use Euripides' *Medea* and St. Paul in his *Letter to the Romans* as standard examples of implementation failure.

Medea's desire to harm her unfaithful husband, Jason, brings about a relatively unforgettable case of implementation failure. In the dramatic climax of the play, her desire overcomes her careful reasoning and leads her to murder her own children in order to avenge herself on her husband (Euripides, 2008). In Ovid's telling of the story, she laments, "I am dragged along by a strange new force. Desire and reason are pulling in different directions. I see the right way and approve it, but follow the worse" (Ovid, 2009). Somewhat more moderately, St. Paul's comments in the seventh chapter of his Letter to the Romans also serve as a standard illustration of implementation failure. Describing his inability to carry out his many positive intentions, he writes: "We know that the law is spiritual; but I am unspiritual, sold as a slave to sin. I do not understand what I do. For what I want to do I do not do, but what I hate I do. And if I do what I do not want to do, I agree that the law is good" (Romans 7:14-17). In this way, both Medea and St. Paul represent paradigm cases of what I will call implementation failure.

Distinguishing between implementation failure and explanations of weakness of will more easily allows us to answer the first issue outlined above, namely, 'Do instances of weakness of will genuinely exist?' Understood as referring to the behavior of implementation failure, the answer to this question is indisputably 'Yes.' We all do things we know we shouldn't do or hadn't wanted to do, and this position will be held throughout the dissertation. The question that has really challenged philosophers and remains the guiding concern of this project is why we perform these strange kinds of actions and, specifically, what mechanism(s) underlie them.

3. Two Philosophical Models

Throughout the history of the tradition, philosophers have mainly sought to discover the causes – or mechanisms - underlying weakness of will. In Part I of this dissertation, I discuss two main schemas for doing so, namely, by offering either a 'syllogism-based' or 'valuation-based' models of weakness of will. The logical form of the syllogism provides the conceptual structure for the first of the two mechanisms. 'Valuation,' or the processes whereby we come to value and seek out what benefits us as living organisms and avoid what is detrimental, serves as the guiding principle for the second of the two. Syllogism-based models of weakness of will typically specify the following three principles:

Structure	That accounts of practical reasoning rely on deductive or inductive syllogisms
Conflict	That weakness of will arises out of a situation of conflict involving two syllogisms whose contradictory conclusions have opposite truth values and, finally,
Arbitration	That an affective faculty or an autonomous faculty of the will mediates between these competing syllogisms.

By contrast, valuation-based models of weakness of will typically stipulate:

Valence	That agents attribute values to internal and external objects and events
Activation	That positively valuated objects and events elicit approach responses, while negatively valuated objects and events elicit withdrawal responses and, finally,
Error	That error is the product of agents evaluating an alternative as *apparently* more valuable than it actually is

Syllogism-based models have systematically dominated philosophical treatments of the issue throughout the history of philosophy, and remain prevalent in contemporary philosophy (Harman *et al.* 2011). One avenue to reinvigorate this overly narrow debate is by highlighting

the less known, and arguably more naturalistically compatible, valuation-based models of weakness of will.

Chapter 2 analyzes Aristotle's account of weakness of will and shows how it became the prevailing philosophical treatment of the issue throughout much of medieval and modern philosophy. It further examines how Donald Davidson reintroduces a quasi-Aristotelian theory of action and weakness of will, reconstructing Davidson's analysis of weakness of will in his 1970 essay, "How is Weakness of the Will Possible?," and discusses Davidson's substantial influence on contemporary theories of weakness of will.

Chapter 3 outlines an alternate, valuation-based model of weakness of will. It analyzes the positions of Plato's Socrates (in both the *Protagoras* and *Republic*), Spinoza, and Hare and argues that, although they clearly set out from different philosophical points of departure, they share in common the views that: a) an agent attributes values to internal and external objects and events, b) positively valuated objects and events elicit approach responses, while negatively valuated objects and events elicit withdrawal responses and, finally, c) although an agent must knowingly evaluate something as apparently more valuable in order to pursue it, this goal may nonetheless be objectively less valuable than the alternatives.

The two models are then evaluated using Craver and Darden's criteria for assessing mechanism schemas (see Section 5 of this chapter, below).

4. A New Mechanistic Account of Weakness of Will

Part II of the dissertation continues in this historical tradition by searching for the mechanism underlying weakness of will. In particular, it aims to develop a mechanism

schema of weakness of will that that is consistent with contemporary behavioral psychology and computational neuroscience.

Broadly, converging evidence from computer science, psychology, and neuroscience indicates that the human brain employs three dissociable mechanisms to make choices. The 'Pavlovian' mechanism corresponds to 'hard-wired' approach and withdrawal responses. 'Goal-directed' behaviors map out different options and assess them in light of specific goals. 'Habit-based' behaviors learn the value of actions over time and in a given situation choose the most consistently valuable option in that situation. Although weakness of will is traditional identified as a single phenomenon, I argue that suboptimal interactions between these three decision-making mechanisms generate two different categories of weakness of will, which are etiologically but not psychologically distinguishable.

To make this clear, Chapter 4 considers how reinforcement learning in computer science investigates optimal decision-making systems, focusing on the computational models that characterize the Pavlovian, goal-directed, and habit-based decision-making mechanisms. I contend that weakness of will in fact consists of a suite of discrete behaviors that include 'habit-based' and 'Pavlovian' categories of weakness of will. In 'habit-based' weakness of will, agents rely on familiar actions to navigate everyday situations, only to realize in some cases that the circumstances have changed and their actions are no longer appropriate. In 'Pavlovian' weakness of will, agents' recognize the best course of action, but 'hard-wired' responses limit their ability to contemplate or pursue alternative courses of action.

5. Evaluating Competing Mechanisms

The search for mechanisms involves committing to what Craver and Darden call a 'garden-variety' realism that assumes that full-fledged target mechanisms such as DNA repair

mechanisms or cellular transport mechanisms really exist (2013, 9, 68).[5] Subscribing to this kind of moderate realism in turn requires ways to evaluate and test competing mechanism schemas.

One obvious way to do this involves experimentation. As Bechtel and Abrahamsen observe, "a researcher tests hypothesized mechanisms by inferring how the mechanism or its components will behave under specified conditions and uses the results of actually subjecting the system to these conditions to evaluate the proposed mechanism" (2005, 436). This would mean, for instance, that in the example of a broken doll, a child forms a hypothesis about how the limb fits into the socket, and tries to reattach it accordingly. Her hypothesis is then corroborated or disconfirmed based on whether the limb can be put back into the socket or not.[6] Along similar lines, more complex experiments are carried out at all levels of investigating inorganic and organic mechanisms, including chemical reactions, transport mechanisms, regulations mechanisms, reproduction mechanisms, and so forth.

Physical experiments are not the only means available for evaluating mechanisms, however – nor, arguably, should they be the first line of defense. Mechanisms can and ought to be assessed theoretically, before the costs of experimentation are taken on. Craver and Darden delineate three criteria for evaluating mechanisms, concentrating on the ways in which mechanisms can succeed, fall short or fail altogether.

The first class of mechanism schema failure is 'superficiality' (Craver and Darden, 2013, 87-89; see also Craver 2006). Superficial schemas re-describe a phenomenon without providing an account of an underlying or internal mechanism that would generate it. They

[5] Or at least 'exists' in some sense. Craver and Darden note, "one can acknowledge the ideals… for descriptions of mechanisms while, at the same time, recognizing that science traffics in idealized and incomplete schemas" (2013, 9).

[6] Conditions change, however: the child may understand how the socket should work, but the socket can have become damaged. Reattaching the limb would then require a different strategy, such as using an elastic band.

are also sometimes called 'phenomenal explanations.' For example, if one were to ask, 'How does a car engine work,' the response, 'When I turn on the key, the car starts' would constitute a superficial mechanism schema. Other standard cases of superficial schemas involve elements such as phlogiston, at one time thought to be responsible for combustion, or the use of homunculi to explain different behaviors (Craver and Darden 2013, 87).

The second class of failure is 'incompleteness' (Craver and Darden, 2013, 89-94). Incomplete mechanism schemas are mainly characterized by the fact that they use placeholders where the relative sub-mechanisms involved in producing a phenomenon are not yet fully understood. "Sometimes gaps are marked in visual diagrams by black boxes or question marks," Craver and Darden specify, while other times they are "masked by *filler terms*. Terms such as activate, cause, encode, inhibit, produce, process, and represent are often used to indicate a kind of activity in a mechanism without providing any detail about how that activity is carried out" (2013, 31).

Effective schemas typically minimize incompleteness over time; that is, discovery moves from 'black box sketches,' for which neither component nor functions are determined, to sketches containing 'gray boxes,' for which component functions are identified, to schemas, which are comprised of 'glass boxes' (2013, 31). Glass boxes improve on grey boxes by completely and accurately describing the functional sub-component in question. Craver and Darden describe this as moving from 'mechanism sketches' to 'mechanism schemas.' The former correspond to incomplete representations, where some parts, activities, and organizational features are specified, but there remain explanatory gaps. If a sketch turns out to be on the right track, it can develop into a schema. The latter correspond to a "description of a mechanism, the entities, activities, and organizational features of which are known in sufficient detail that the placeholders in the schema can be

filled in as needed" (2013, 31). Another way of thinking of mechanism schemas is that they are "complete enough for the purposes at hand" (2013, 31). A good example of an incomplete mechanism schema would be the laws of Mendelian inheritance. The original Mendelian account contains a 'black box' when it comes to the carriers of genetic material. The same box is subsequently made into a gray box by the Boveri-Sutton theory, which identifies chromosomes as the genetic carriers.

Finally, the third class of schema failure is 'incorrectness' (Craver and Darden 2013, 94-95). Here, the schema is evaluated based on whether one or more of its components are corroborated by, compatible with or explicitly at odds with existing empirical evidence. As with incompleteness, incorrectness is a matter of degree. For example, it may be possible that dogs' stomachs digest grass in a certain way, that there is no existing evidence regarding dogs digesting grass, or that there is evidence that dogs' stomachs fail to digest grass. Ideally, a mechanism achieves the status of a 'how-actually' schema when it describes how the mechanism "in fact works," in this case, accurately describing how enzymes break grass down into chyme (Craver and Darden 2013, 94; example mine).[7] Again, this characterization must be understood in terms of the 'garden-variety' realism, where the schema is considered to be "correct enough rather than correct full stop" (*Ibid.* 95).

[7] (In fact, there is limited evidence indicating that the canine digestive tract cannot adequately digest grass or other non-grass plants (Sueda *et al.* 2007)

Target Mechanism	**s → a → b → c→ d → f**
A Non-Superficial, Complete, and Correct Schema	S → A → B → C→ D → F
A Superficial Schema	S → F
An Incomplete Schema	S → D → [redacted] → F
An Incorrect Schema	S → P → Q → R → F

Figure 1.2. Superficial, incomplete, and incorrect schemas. Top row: 's' represents the 'starting conditions' of the target mechanism. 'f' represents the 'finishing conditions' of the target mechanism. 'a,' 'b,' 'c,' 'd,' represent the functional subcomponents of the target mechanism. Bottom rows: 'S' represents the 'starting conditions' in the mechanism schemas. 'F' represents the 'finishing conditions' in the mechanism schemas. The letters 'A,' 'A,' 'C,' 'D,' 'P,' 'Q,' and 'R,' represent the functional subcomponents of the mechanism schemas. (Reproduced from Craver and Darden 2013, Figure 6.1)

Throughout the dissertation, Craver and Darden's three criteria will be used at the end of each relevant chapter to evaluate the competing traditional philosophical models of weakness of will (Figure 1.2.).

In Chapter 5, I use Craver and Darden's (2013) criteria to evaluate my own account of weakness of will. I particularly focus on what they call the 'vices' of incompleteness and incorrectness and argue that although my mechanism schema of weakness of will is not complete, it is correct, i.e., it is consistent with current scientific evidence.

Having arrived at a preliminary but correct mechanism schema for weakness of will, in Chapter 6, I describe how this dissertation evolved from indifferently devising a counterexample to a philosophical argument to seriously searching for the mechanisms underlying weakness of will. I conclude the dissertation by discussing the central implications of my research.

CHAPTER 2

THE STANDARD THEORY: SYLLOGISM-BASED EXPLANATIONS OF WEAKNESS OF WILL

1. Introduction

Nearly every major philosopher from Aristotle to Donald Davidson has offered a detailed account of whether and how weakness of will is possible. Nevertheless, the issue continues to be the subject of a surprisingly lively debate in contemporary philosophy. Following Davidson's influential account in "How Is Weakness of the Will Possible?" (1970), authors including Michael Bratman (1979), Alasdair MacIntyre (1990), Richard Holton (1999), and Nomi Arpaly (2000) have brought forward different explanations of how weakness of will can be logically consistent, and argued for weakness of will as both a rational and irrational behavior. Dozens of articles and books continue to come out on the topic, with two recent volumes, Alfred Mele's *Backsliding: Understanding Weakness of Will*, published as recently as April 2012.

Over the course of this long interpretive tradition, philosophers have presented two types of mechanisms to try and explain weakness of will. I call these two types of mechanism schemas the 'syllogism-based' and 'valuation-based' models of weakness of will. As the name of the former account suggests, the syllogism-based models are characterized by the rules of formal logic. In general, this type of mechanistic explanation of weakness of will specifies:

Structure	Deductive or inductive syllogism-based accounts of practical reasoning
Conflict	That weakness of will arises out of a situation of conflict involving two syllogisms whose contradictory conclusions have opposite truth values and, finally,
Arbitration	That an affective faculty or an autonomous faculty of the will mediates between these competing syllogisms.

In this chapter, I argue that syllogism-based accounts have historically dominated philosophical debates of practical reasoning and weakness of will but are, on Craver and Darden's criteria, incomplete and incorrect

In Section 2 of this chapter, I provide a brief account of practical reason as understood within the framework of the logical syllogism. In Section 3, I discuss Aristotle's account of weakness of will and show how it became the prevailing philosophical treatment of the issue throughout much of medieval and modern philosophy. I show that the positions of Aristotle, Aquinas, and Descartes share in common: a) strictly syllogism-based models of practical reasoning, where weakness of will arises out of a situation of conflict involving two equal but contradictory syllogisms b) the view that these competing syllogisms are mediated by an affective faculty or an autonomous faculty of the will.

In Section 4, I examine how, in responding to Hare and the 'logical connection argument,' Donald Davidson reintroduces a quasi-Aristotelian theory of action and weakness of will. I reconstruct Davidson's analysis of weakness of will in his 1970 article, "How is Weakness of the Will Possible?" I then outline Davidson's substantial influence on contemporary theories of weakness of will. In particular, I highlight internalist and external approaches to 'unconditional' weakness of will.

In Section 5, I turn to Craver and Darden's criteria for evaluating mechanisms and explain how Aristotle, Davidson and 'Davidsonian' accounts suffer from the 'vices' of incompleteness and incorrectness. I conclude by suggesting that syllogism-based models of practical reasoning may struggle to explain weakness of will because they lack a functional sub-mechanism responsible for evaluating better and worse alternatives.

But let us start by looking at practical reasoning, broadly construed.

2. Practical Reasoning and Deductively Valid Arguments

The term 'practical reason' refers to the human ability to decide what to do. In contrast to the notion of 'theoretical reason,' which is the ability to decide what to believe, practical reason is oriented towards action in both its subject matter and in its product, with the process of practical reasoning thought to result in either a judgment regarding an action or in the execution of one. The difference between these two kinds of reasoning can be seen in the way that they are expressed: theoretical reasoning is usually expressed in the form of declarative statements, such as, 'It is raining outside,' while practical reasoning is usually expressed using imperatives, such as, 'We should stay inside,' or by physically 'staying inside.'[8]

The prevailing model of practical reasoning follows the structure of a logical syllogism and, in particular, of a valid deductive syllogism. On this account, a person's reasoning might start out with a general rule, proceed to a specific premise regarding the situation at hand, and conclude with either a judgment regarding action or the action itself. For example, when we are deciding whether or not to go outside in the rain, a deductive model could represent our reasoning process in one of the following two ways, either as an argument diagram or as a syllogism (Figure 2.1.).

[8] Of course, in everyday life, the two kinds of reasoning are often combined, e.g.: 'It is raining, so we should stay inside.'

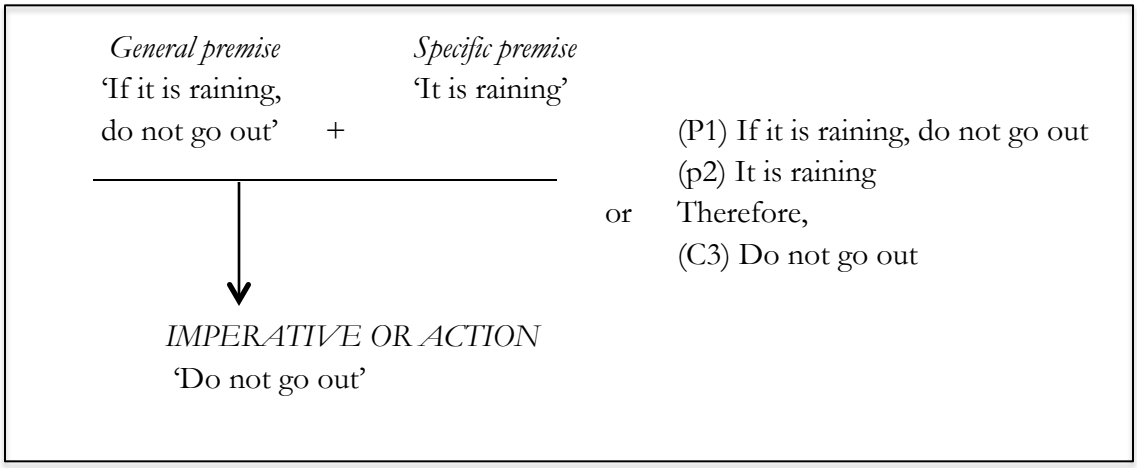

Figure 2.1. On the prevailing model of practical reasoning, deliberation can be represented in the form of a valid deductive syllogism. This deductive syllogism can itself be characterized using either an argument diagram or a standard three-part deductive syllogism.

In addition, most proponents of the syllogistic model of practical reasoning assume or even express some combination of the following three claims:

a) Deductively valid arguments make people justified in believing the conclusions of those arguments[9],
b) A person's beliefs in the premises cause that person to believe the conclusion of a deductively valid syllogism, and
c) The premises are independent; the universal premise (P1) contains all of the moral content, and the other premises are morally neutral (Harman et al., 2011, 214).

Claims (a) and (b) in particular raise the fundamental question of *how* reasons can be translated into and/or cause actions, and explain why weakness of will poses such a significant explanatory obstacle. That is, if people subscribe to the premises of an obviously valid argument, how can they fail to accept and/or carry out the conclusion? For syllogistic theories of practical reasoning, the phenomenon of weakness of will poses an explanatory

[9] In other words, proponents of the syllogistic model of practical reasoning assume that there is a connection between the principles of logic and the principles of reasoning. The philosopher Gilbert Harman calls this the 'Logical Implication Principle,' (LIP) which states, "the fact that one's view logically implies *P* can be a reason to accept *P*" (1988, 11; see also Harman 2002). Harman wants to criticize the Logical Implication Principle and draw a strong distinction between logical inference and reasoning. I think the issue of 'soundness' (i.e. the question of true premises together with a valid argument) moves us into the territory of 'beliefs,' and hence, 'reasoning.'

obstacle; it is a feature of our everyday lives but it is also genuinely counter-intuitive, and seems to defy many of our assumptions regarding rational human activity. Accordingly, most syllogistic theories of practical rationality have had to take up the challenge of accounting for the nature and underlying causes of weakness of will.

3. Aristotle and the Practical Syllogism

This section focuses on what is evidently the earliest syllogism-based model of practical reasoning, namely that of Aristotle, paying particular attention to Aristotle's mechanism for explaining weakness of will. I then go on to provide a brief, supplementary sketch of the reception of Aristotle's account of weakness of will. Aristotle's main discussion of weakness of the will is presented in *Nichomachean Ethics* Book VII, Chapter 3. I turn to Davidson's discussion of weakness of will in Section 4.

3.1. How Weakness of Will Works for Aristotle

Aristotle's main discussion of practical reasoning and weakness of will is presented in *Nichomachean Ethics* Book VII, Chapter 3. His account is set out in four parts. The first three parts show how practical reasoning generally proceeds. The fourth part takes up these preceding elements to provide an explanation of how weakness of will occurs (Santas, 1969, 181).

In the first passage (1146b31-36, Translated by W.D. Ross, revised J.O. Urmson, in Barnes 1995, Volume 2), Aristotle distinguishes between two senses of what it means 'to know.' In the first sense, an individual both has and exercises her knowledge about something. In the second, she has this knowledge but doesn't use it. This distinction is decisive for cases of weakness of will, since "when a man does wrong it will make a

difference whether he is not exercising the knowledge he has, (viz., that it is wrong to do what he is doing), or whether he is exercising it" (1146b33-34). As Dahl (1984, 14) puts it, the former case would be unremarkable, but the latter would be "strange."

In the second passage (1146b36-1147a9), Aristotle discusses the practical syllogism and specifies that knowledge consists in two further sub-components, namely, a universal premise and a particular premise. For example, an individual may know the universal premises 'dry food is good for all men' and 'toast is dry food,' and she may also know the particular premise, 'this object in front of me is toast.' An akratic individual, however, may only know 'dry food is good for all men' and 'toast is dry food,' while not knowing or not exercising her knowledge that the object in front of her is toast. As such, Aristotle notes, "it may well happen that a man knows both [the] major and minor premise of a syllogism and yet acts against his knowledge, because the minor premise which he uses is universal rather than particular. In that case, he cannot apply his knowledge to his action, for the actions to be performed are particulars" (1147a1-4). And once again, this distinction is essential for considering cases of weakness of will, since 'having knowledge' of the universal but *not* the particular premise would be perfectly understandable, but, as Aristotle remarks, "it would be surprising if he 'knew' in the other sense, namely with both terms apprehended as concrete particulars" (1147a9), that is to say, it would be strange if the agent knew both the universal premise and recognized the particular situation to which it applied, but did not act accordingly.

In the third passage (1147a10-24), Aristotle discusses a special set of cases in which an individual may have knowledge but is not exercising it. He includes examples of when an individual is "asleep, mad, or drunk," but also specifies that this is precisely the situation of

individuals when they are "in the grip of the emotions," and of incontinent individuals (1147a15).

Finally, in the fourth passage (1147a24-b19), Aristotle gathers up elements from the preceding three passages and offers an explanation of what happens in a moment of weakness of will. He presents his explanation in two sub-sections. In the first sub-section, he describes a case in which everything goes 'according to plan.' He explains, "the one opinion is universal, the other is concerned with the particular facts, and here we come to something within the sphere of perception; when a single opinion results from the two, the soul must in one type of case affirm the conclusion, while in the case of opinions concerned with production it must immediately act" (1147a25-28). In this way, when both the universal premise and the particular premise combine, they produce either the correct inference, or even more directly, an action (Figure 2.2.). For example, if an agent holds that 'everything sweet ought to be tasted,' and recognizes that a certain object is sweet, then "the man who can act and is not restrained must at the same time actually act accordingly" (1147a31).

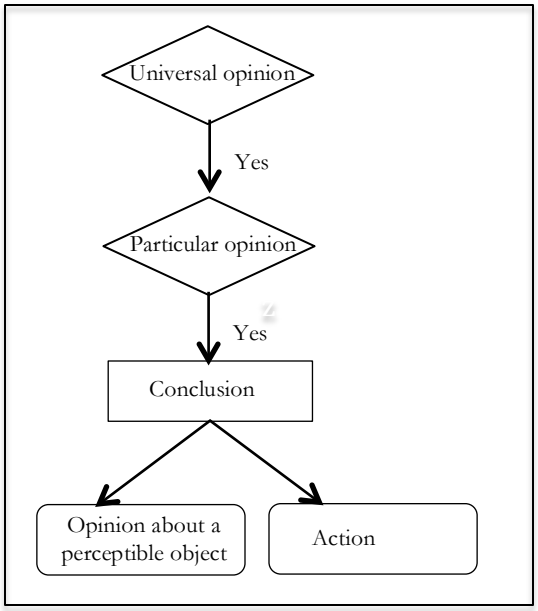

Figure 2.2. When both the universal premise and the particular premise correspond to produce either the correct inference, or even more directly, an action.

In the second sub-section, Aristotle presents an account of what happens when something "goes wrong," (Santas, 1969, 183) or in other words, what takes place in an instance of weak-willed action. Here, the agent holds two universal opinions, such as that 'unhealthy things ought to be avoided' and 'everything sweet is pleasant,' and at the same time holds the particular belief that a certain object in front of her is sweet. This would result in conflicting conclusions. This situation can be represented in two parallel syllogisms (Figure 2.3.).

($P1^1$) All unhealthy things ought to be avoided	($P1^2$) All pleasant things ought to be eaten
($P2^1$) All sweet things are unhealthy	($P2^2$) All sweet things are pleasant
($p3^1$) This is a sweet thing	($p3^2$) This is a sweet thing
Therefore,	Therefore,
($C3^1$) This sweet thing ought to be avoided	$C3^2$) This sweet thing ought to be eaten

Figure 2.3. Agents occasionally face decision-making dilemmas. On Aristotle's account, these kinds of dilemmas can be represented in the form of two competing syllogisms, where each syllogism which issue an opposite appropriate action.

Aristotle specifies that if the particular premise ('This is a sweet thing') is "active", *and* the agent has appetite, then she will eat the object in front of her). But what exactly happens here? If 'this is a sweet thing' is active, why does it not combine with both $P1^1$ to produce avoidance *and* $P1^2$ to produce approach? In other words, why does it only combine with $P1^2$ and $P2^2$ to produce eating?

At first, one is tempted to look to what Aristotle has to say earlier, about "having knowledge but not using it" (1984, 1812/1147a11-12). Building on his preceding analysis, Aristotle observes, "Now, the last proposition both being an opinion about a perceptible object, and being what determines our actions, this a man either has not when he is in the state of passion, or has it in the sense in which having knowledge did not mean knowing but only talking, as a drunken man may utter the verses of Empedocles" (1984, 1812/1147b9-

12). In the example of the sweet thing to be either avoided or eaten, this 'last proposition' would correspond to 'This is a sweet thing.'[10] But as Santas (1969, 183) points out, 'This is a sweet thing' plays a role in both sets of opinions, and "clearly, if it were to turn out that what goes wrong in the ['To be avoided' set of opinions] implies that the same thing goes wrong in the ['Eat this' set of opinions], we would be left without an explanation of the action at all. So this difficulty has to be overcome if Aristotle's view on incontinence is to be coherent at all." Instead, following Vlastos (1966), Santas offers the following explanation, (1969, 183).

Santas' solution to the problem hinges on Vlastos's suggestion that *he teleutai protasis* ('the last proposition') does not necessarily refer to the last *premise* ('This is a sweet thing,'), which is shared by both sets of opinions, but may refer instead to the conclusion, 'This sweet thing ought to be avoided.' In this way, what the weak-willed individual knows but does not exercise is the *conclusion*, 'I should not eat this cake.' This avoids the problem of the shared premise discussed above, and also coheres well with Aristotle's suggestion that "because the last term is not universal nor equally an object of knowledge with the universal term, the position that Socrates sought to establish actually seems to result; for it is not what is thought to be knowledge proper that the passion overcomes (nor is it this that is dragged about as a result of the passion), but perceptual knowledge" (1984, 1813/1147b14-17). On this interpretation, Aristotle can claim to both align his account with Socrates' view in the *Protagoras*, as well as to offer an explanation of weakness of will that does not appear to 'contradict the phenomena.'

In this way, Aristotle specifies that the weak-willed individual has knowledge of both the particular and universal, and even infers the appropriate conclusion, but does not exercise this knowledge because it is 'overcome by passion' (1984, 1813/1147b17). Aristotle

[10] Saarinen (1994, 12) draws on the *Prior Analytics* to suggest that both the major and the minor premise are known, but somehow the individual fails to connect them.

thus exemplifies the third feature typical of syllogism-based accounts, namely, that weakness of will occurs when an affective faculty or an autonomous faculty of the will mediates between two competing syllogisms.[11]

At the same time, Aristotle's account thus still leaves open the question, 'What exactly does it mean for the conclusion to be 'overcome by passion?' Or in other words, in terms of a mechanistic account, one is left wondering, 'How exactly would this work?' I return to this issue in Section 6, where I suggest that Aristotle's account of weakness of will is incomplete and incorrect.

3.2. Aristotle's Influence on Medieval Philosophy

Despite some of its shortcomings, Aristotle's model of weakness of will appeared to provide a fundamental moral psychological account and remained highly influential well into the medieval period. In the end of the fourth and beginning of the fifth centuries AD, Augustine continues to conceive of weakness of will as the product of inner mental turmoil brought about by competing reasons and desires.[12] For his part, Aquinas has renewed access to Aristotle's *Nichomachean Ethics,*[13] and follows Aristotle to argue that weakness of will is caused by desires corrupting the practical syllogism. Aquinas defends this view in both his commentary on Aristotle's *Nichomachean Ethics* as well as in his own original positions

---

[11] I will not attempt further analysis of the scholarly debate here, but rather accept Santas' position as a plausible reading of Aristotle, and an intuitively plausible account of weakness of will.

[12] Augustine, *Confessions*, John K. Ryan (Trans.), 1960, 8.5.11-12; 'The Deserts and Remission of Sin', in *Augustine: Later Works*, Ed. Burnaby, 1955, 187. Augustine is widely held to have introduced the concept of the will into the Western canon. But this is a misconception introduced by Albrecht Dihle in his influential work, *The Theory of Will in Classical Antiquity*. Here, Dihle argued that "the notion of the will, as it is used as a tool of analysis [...] from the early Scholastics to Schopenhauer and Nietzsche, was invented by Augustine" (1982, 123). However, while acknowledging that Dihle has been highly and rather unfortunately influential, the majority of contemporary scholars of medieval philosophy agree that it is deeply unlikely that Augustine would have understood the will as a separate psychological faculty. Rather, they suggest that he would conceived of the will as a mixed mental power, much like memory, to which he explicitly compared it in *On Free Choice* (*De libero arbitrio,* 2.19.51). For an extended critique of Dihle's analysis, see O'Daly, 1989; Saarinen, 1994).

[13] Aquinas has access to it in the form of Robert Grosseteste's translation from 1247. Augustine does not have access to the majority of the Nichomachean Ethics (Saarinen, 1994)

developed in his *Summa Theologica* and in *De Malo*.[14] In this way, the framework for interpreting weakness of will as a conflict between one's beliefs and desires spans an interpretive period of nearly seventeen centuries, and continues well into the thirteenth century.

A second interpretive period for syllogism-based theories of weakness of will is brought about quite abruptly. Broadly, on this view conflicts between opposing syllogisms are no longer mediated by affects or the emotions, but rather by an autonomous faculty of the will.

In the late thirteenth century, the Paris Condemnation of March 1277 forces theologians and philosophers to adopt strongly voluntarist principles, that is, principles that attributed a high degree of autonomy and power to the human will.[15] As a consequence, the traditional model of the mental faculties comes to include a will that is understood to have absolute freedom and authority in determining human action. As one possible author of the Condemnation, Walter of Bruges, formulates it, this new relationship between the mental faculties can be understood in the following way: "the intellect moves the will as a counselor moves the pope [...] – not as an efficient cause, not as a great power that impels or necessitates, but by persuasion, by presenting the good."[16] This means the will is not determined either by desires (even strong desires) or by the intellect; it has absolute freedom. Naturally, in keeping with these principles, philosophers are also compelled to revise the 'belief and desire' model of weakness of will. As we see in the writings of Henry of Ghent,

---

[14] Of the *Sententia libra Ethicorum*, Saarinen (1994, 118) writes that, "in the commentary Thomas often follows Aristotle's text closely and does not aim at an original contribution," emphasizing that the more substantial discussion of his views on *weakness of will* are presented in the *Summa Theologica* (1-2 Q77A1-A2, 2-2 Q156 A1) and in *De Malo* (Q3 A9).

[15] The Condemnation of March 1277 prohibits the teaching of 219 philosophical and theological theses, many of them Averroist, and includes an order to teach a strongly voluntarist conception of free will (Kent, 1995, 69).

[16] Cited in Kent, 1995, 119-20.

Walter of Bruges, and to a lesser extent, John Buridan, weakness of will is no longer understood as the product of a conflict between reasons and desires, but is rather thought to be caused by the autonomous movement of the will. For example, revising Aquinas' account, Walter of Bruges substitutes for desires a free and sovereign will, so that it is now the will which deliberately frustrates the workings of the practical syllogism.[17]

In many respects, this new understanding of the mental faculties makes the problem of weakness of will relatively easy to resolve, since only a sinful will is needed to explain it. And indeed, philosophers and theologians continue to embrace the will-based model of weakness of will throughout the later medieval period. In the seventeenth century, we see Descartes struggle to clarify his understanding of the relationship between a free will and weak-willed action; he ultimately interprets the weakness of will phenomenon in terms of a corrupt free will.[18]A modified version of this account still characterizes Kant's eighteenth century understanding of weakness of will (Hill Jr. 2012).

4. Davidson's Rejection of the Deductive Account of Practical Deliberation

In the second half of the twentieth century, Donald Davidson presents his model of weakness of will as an improved version of Aristotle's classical account (1980, 32). Retaining several features of Aristotle's interpretation, he nevertheless rejects the requirement of a deductively valid syllogism, and proposes an inferential account similar to Hempel's analysis of probabilistic evidence (Hempel, 1965).[19] I suggest that many of Davidson's amendments

[17] *Quaestiones disputatae du B. Gauthier de Bruges: texte* inedit, Longpre (Trans.), 1928. See also, Saarinen, 2011, 32.
[18]Descartes discusses weakness of will in at least four texts, including in Part III of the *Discourses*, in a letter to Mersenne (1637), in two letters to Mesland (1644, 1645), and in his last work, *Passions of the Soul* (1649); see also Alanen, 2003; Ong-Van-Cung, 2003; Pironet and Tappolet, 2003.
[19] That is, Davidson proposed to substitute the traditional, deductive account, where it is impossible for the premises to be true and the conclusion false, to an inductive model based on probability.

to the classical account emerge in response to his contemporaries' denial of the view that reasons can provide causal explanations for actions.

4.1. Reasons as Causes

In the mid-twentieth century, Aristotle's understanding of the practical syllogism was the subject of much critique.[20] In what became known as the 'logical connection argument,' Wittgenstein, Melden, and others argued that since a) causes must be logically distinct from their effects, and b) reasons and actions are logically (deductively) interconnected, then c) reasons cannot provide causal explanations of actions. Philosophers broadly agreed that the deductive nature of the successful practical syllogism precluded it from serving as a causal relation between reasons and actions.[21] However, in his 1963 article "Actions, Reasons and Causes," Davidson broke with this tradition to defend what he called the "ancient – and commonsense – position that rationalization is a species of causal explanation" (1980, 3).

Davidson's account of reasons as causes relies on two key concepts. First, Davidson argues that actions are to be rationalized or explained in terms of 'primary reasons,' where each primary reason consists of a belief and a desire, or as he calls the latter entity, a 'pro-attitude.' A pro-attitude can correspond to "wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values insofar as these can be interpreted as attitudes of an agent directed toward actions of a certain kind" (1980, 4). In this way, the act of flipping on the light switch can be explained in terms of the following primary reason: I *know* flipping the switch turns on the light (belief) and I *wanted* to turn on the light (pro-attitude).

---

[20]Glüer 2011, 158.

[21] Stoutland, 1971, Ryle, 1949, Anscombe, 1959, Hampshire, 1959.

Second, Davidson emphasizes that a single action can be conceived of in several different but equally 'correct' ways. Staying with the same example, for instance, at the moment that I flip the switch, I turn on the light, but I may also alert a prowler or wake up the baby. My act of flipping the light switch can thus be described as the intentional act of turning on the light, and/or as the unintentional act of alerting the prowler or waking the baby. Importantly, however, beliefs, desires and bodily movements or actions can also be understood as *events*. My wanting to turn on the light and my act of flipping the switch can also be described as events, and so the connection between my reason and my action can also be understood *as a connection between two events*. On the basis of this, Davidson argues that the "mysterious connection" between a reason and an action can be simultaneously rationalizing and causal after all: the 'primary reason' rationalizes or explains the action, but it is also causal, insofar as it is described as an event that causes the further event of flipping the switch. In this way, Davidson defends the role of reasons as both explanations and causes.

4.2. Weakness of Will

In the context of his analysis of reasons as explanations *and* causes, Davidson's interest in the problem of weakness of will becomes substantially clearer. Where Socrates conferred on Aristotle the task of explaining the experiential phenomenon of implementation failure, Davidson comes to the issue from the perspective of reasons and their dual role as explanations and causes. In particular, in examining the problem of weakness of will, Davidson must solve the apparent *logical* inconsistency at the center of the issue, i.e., the problem of if reasons are causes, then how is it possible for the strongest reasons not to produce the corresponding action? At the same time, he must defend weakness of will as

causally (physically) possible. As he remarks in the introduction to his collected essays, "if reasons are causes, it is natural to suppose that the strongest reasons are the strongest causes. [However,] I defend the causal view [in "How is Weakness of the Will Possible"] by arguing that a reason that is causally strongest need not be a reason deemed by the actor to provide the strongest (best) grounds for acting" (1980, xii). He proceeds by showing that reasons that serve as *explanations* do not necessarily entail the implementation of corresponding reasons as *causes*. If this is the case, then weakness of will is both logically and causally (physically) possible. But in doing so, Davidson also largely shifts his attention away from the task of explaining the phenomenon of implementation failure, and focuses instead on harmonizing the logical structure of practical reasoning in the context of weakness of will.

Davidson broadly defines weakness of will as follows:

> D: In doing *b* an agent acts incontinently if and only if: (a) the agent does *b* intentionally; (b) the agent believes there is an alternative action *a* open to him; and (c) the agent judges that, all things considered, it would be better to do *a* than *b* (1980, 22).

Very quickly, he also remarks that these kinds of actions "seem" to exist or are even quite "certain" to exist (1980, 22, 29). Nevertheless, he points out that when broken down into detailed principles, the definition of weakness of will seems to contain a fundamental logical impossibility. He breaks his analysis down into three principles. The first principle refers to the apparently natural relationship between desiring or wanting to do something and doing it (action):

> P1: If an agent wants to do *x* more than he wants to do *y* and he believes himself free to do either *x* or *y*, then he will intentionally do *x* if he either does *x* or *y* intentionally (1980, 23).

The second principle pertains to the relationship between evaluative judgments of what it is better to do and wanting to do something:

> P2: If an agent judges that it would be better to do *x* than to do *y*, then he wants to do *x* more than he wants to do *y* (1980, 23).

And put together, P1 and P2 clearly entail that 'if an agent judges that it would be better to do *x* than to do *y*, then he will intentionally do *x* if he either does *x* or *y* intentionally' (1980, 23). In turn, this would seem to suggest that the third and last principle must be false:

> P3: There are incontinent actions (1980, 23).

This is precisely the conclusion that Davidson wants to deny. He argues that there is something compelling and persuasive about the pairing of P1 and P2, something which "has seemed to many philosophers, from Aristotle on, to promise to give an analysis of what it is to act with an intention; to illuminate how we explain an action by giving the reasons the agent had in acting; and to provide the beginning of an account of practical reasoning, i.e., reasoning about what to do, reasoning that leads to action" (1980, 31). At the same time, he wants to defend the possibility and existence of incontinent or weak-willed actions. Correspondingly, he spends the remainder of "How is Weakness of the Will Possible" defending the consistency of the three premises P1-P3.

Preparing the foundations for his own theory, Davidson criticizes Aristotle's account for failing to provide an explanation for how moments of weak-willed action are genuinely possible. On his view, Aristotle strictly conceives of the conclusion of the practical syllogism as an action rather than as a propositional claim, and this in turn results in an overly stringent form of the practical syllogism. Specifically, when Aristotle describes a strong desire on the basis of which the agent acts, then he must also attribute to the agent a strong judgment that the action is desirable. This means that P1 and P2 directly contradict P3, and hence leave no 'room' for weakness of will.

Crediting Aquinas with a better interpretation, Davidson argues that a meaningful explanation of incontinence must account for that desire which actually causes the action that is carried out. For this reason, he believes Aquinas is on the right track when he presents the weak-willed man as faced with two competing syllogisms (Figure 2.4.).

THE SIDE OF REASON	THE SIDE OF LUST
(M1) No fornication is lawful	(M2) Pleasure is to be pursued
(m1) This is an act of fornication	(m2) This act is pleasant
(C1) This act is not lawful	(C2) This act is to be pursued

Figure 2.4. Davidson follows Aquinas' model in representing the weak-willed agent as being faced with two competing syllogisms.

Nevertheless, when the two conclusions are translated into the form of comparative judgments, e.g., (C1) 'It is better not to perform this act than to perform it, and (C2) 'It is better to perform this act than to not perform it,' they still result in an outright contradiction. On the basis of this, Davidson claims we are justified (and in fact, actually obliged) to introduce "a piece of practical reasoning present in moral conflict, and hence in incontinence, which we have so far entirely neglected," namely, what he calls 'the will' or 'conscience' (1980, 36).

It is not exactly clear why Davidson adopts the controversial and much-abused notion of 'the will,' but in any case, he does not employ it in the traditional, robust sense of the term, i.e., as a free and sovereign faculty of the mind, capable of overriding the intellect and the passions.[22] Instead, he uses the phrase to refer to some capacity in the agent to weigh

THE WILL (CONSCIENCE)
(M3) M1 and M2
(m3) m1 and m2
(C3) This action is wrong

Figure 2.5. Davidson reformulates the practical syllogism to combine the two competing syllogisms. (M3) is a combination of M1 and M2. (m3) is a combination of 'this is fornication' and 'this is pleasant.'

[22] See Saarinen 1994.

and decide between several conflicting reasons. As he explains, "it is not enough to know the reasons on each side: he [the agent] must know how they add up" (1980, 36).[23] To this end, Davidson reformulates the practical syllogism to combine the two competing syllogisms (Figure 2.5.). On this view, when the weak-willed agent 'acts against his better judgment,' he is acting in contradiction to (C3), and not (C1) (from above). It follows that an agent's 'better judgment' corresponds to 'all the relevant factors he or she has considered,' and not what Davidson calls "any judgment for the right side (reason, morality, family, country)," i.e. judgments about 'what is right' (1980, 36).

Still, it remains for Davidson to show how the conclusion (C3) could follow from the premises (M3 and m3) that precede it, and how it could be possible for the agent to nevertheless perform the action in question. Correspondingly, Davidson makes his major argumentative move here: he argues that we must stop conceiving of the major premises of the syllogisms as universalized conditionals. Rather, we should recognize that in practical reasoning, these premises are formulated in relation to certain factors or considerations, and hence are best represented as prima facie statements. For example, instead of saying,

P1: *a* is better than *b*,

we should say,

P1[1]: in light of consideration(s) *c, a* is better than *b*.[24]

Further, a similar formulation should be applied to the entire set of considerations that the agent must weigh and evaluate. In other words, the comprehensive argument should take on the following form. The (M1), (m1) and (C1) syllogism is reformulated to read:

> (M6) *pf* (*x* is better than *y*, *x* is refraining from fornication and *y* is an act of fornication)

[23] In a footnote, Davidson adds, "my authority for how they do add up in this case is Aquinas," and includes a reference to *Summa Theologia* Part II, Q7, Article 2, Reply to Objection 4.
[24] This portion of my analysis is indebted to Stroud (2008).

(m6) *a* is a refraining from fornication and *b* is an act of fornication
:. (C6) *pf*(*a* is better than *b*, (M6) and (m6))

The last line can roughly be said to read, 'In light of (M6) and (m6), all things considered, *a* is better than *b*. *Similarly,* the (M2), (m2), and (C2) and (M3), (m3), and (C3) syllogisms are rewritten to conclude:

(C7) *pf*(*b* is better than *a*, (M7) and (m7))

and,

(C8) *pf*(*a* is better than *b*, *e*), where '*e'* represents the relevant evidence available.

Davidson calls (C8) the "all things considered" conclusion; Stroud (2008) describes it using the analogy of the detective Hercule Poirot, who collects bits of evidence and then comes to a point where he needs to weigh them all against one another. What is essential to Davidson's analysis is that, in contrast to his interpretation of Aristotle, the conclusion drawn here is purely *theoretical*, rather than being an action: "reasoning that stops at conditional judgments such as (C8) is practical only in its subject, not in its issue" (1980, 39). By contrast, only unconditional judgments result in intentional action. As a result, there is no necessary logical relationship between the conclusion of the relational practical syllogism and the action the agent ultimately carries out. The agent can conclude that, all things considered, '*a* is better than *b*,' but still do *b*. Or as Stroud puts it, the inconsistency is equivalent to Poirot thinking "'all the evidence I have seen points toward Colonel Mustard as the guilty party,' [where furthermore] to make this observation is manifestly *not* [for Poirot] to conclude that Mustard is guilty" (2008).

So what causes the agent to perform *b*? Given everything that has been leading up to this moment, Davidson's account is surprisingly brief. He simply remarks that the agent does arrive at an unconditional judgment regarding *b*, and hence does *b*. He concludes, "now

there is no (logical) difficulty in the fact of incontinence […] the logical difficulty has vanished because a judgment that *a* is better than *b*, all things considered, is a relational, or *pf*, judgment, and so cannot conflict logically with any unconditional judgment" (1980, 39). Correspondingly, he rephrases his original definition of incontinence (D) to state that "the agent has a better reason for doing something else," so that "he does *x* for a reason, *r*, but he has a reason r^1 that includes *r* and more, on the basis of which he judges some alternative *y* to be better than *x*" (1980, 40). The fault of incontinence does not involve a logical inconsistency, but rather is a failure in *rationality*, since we consider it rational for an individual to act in accordance with his own best judgment.

From the perspective of preserving the logical consistency of practical reasoning, Davidson's account has been recognized as a successful philosophical theory. But does his account provide a satisfactory explanation of why the phenomenon of weakness of will occurs? I argue that it does not. While acknowledging that the *akrates* performs the problematic action for *some* reason, he cannot account for why this reason prevails over what is otherwise acknowledged to be a *better* reason. Recognizing this, Davidson concludes that, above all, "what is special in incontinence is that the actor cannot understand himself: he recognizes, in his own intentional behavior, something essentially surd" (1980, 42). In this way, on Davidson's account, the fundamental mechanism underlying the conflict between these asymmetrical reasons remains somewhat mysterious. In Section 5, I go on to argue that Davidson's position is not only incomplete but also incorrect.

4.3. Davidson's Influence on the Contemporary Debate

Donald Davidson has been called one of the most influential philosophers in the twentieth century and indeed, his essay "How Is Weakness of The Will Possible?" has had a

tremendous impact on philosophical perspective on weakness of will. At the time of the essay's publication, R.M. Hare was the only major twentieth century philosopher to have published on the topic (1952, 1963). Over thirty major authors have published on the topic since. In many respects, Davidson's original treatment of the issue reintroduced the topic into contemporary analytic circulation.

It is not the aim of this section to provide a detailed account of the post-Davidsonian treatments of weakness of will (see Stroud 2014 for an excellent review). Even a basic sketch endorses the view that syllogistic models of weakness of will continue to dominate contemporary Anglo-American philosophical debates. Following Davidson, philosophers particularly focus on defending the logical possibility of weakness of will. Many philosophers are motivated by the fact that they feel Davidson's account does not go far enough in accounting for weakness of will. In particular, Davidson only guarantees the logical possibility of weakness of will in cases involving 'all things considered' judgments, but not cases of full-fledged, unconditional judgments. This apparent limitation has prompted the next generation of philosophers, among them Michael Bratman, Sarah Buss, and Alfred Mele, to try and establish the logical possibility of full-fledged weakness of will. In doing so, these authors have broadly adopted either 'internalist' or 'externalist' views of motivation. A proponent of internalism holds that an individual cannot make a bona fide judgment about what it would be best to do in a given situation without at the same time – and as a direct result – be motivated to act according to it. In other words, an internalist believes that judgments are themselves intrinsically motivating. By contrast, proponents of externalism argue that an individual can make a judgment about what to do without acting on it, that is, they believe that something else is needed in addition to a judgment in order to motivate an individual to act. They often propose that judgments must be paired with emotions in order

for an agent to be motivated to action (for a good overview of this issue, see Prinz and Nichols 2011, 111-114).

For example, Michael Bratman defends unconditional weakness of will using a thought experiment featuring an individual named Sam. Sam is fully aware of the fact that he should go to bed early in order to wake up the next day, but he is nevertheless in the middle of drinking a whole bottle of wine. A friend of Sam's stops by and, seeing the situation, remarks "'Look here. Your reasons for abstaining seem clearly stronger than your reasons for drinking. So how can you have thought that it would be best to drink?'" (Bratman 1979, 156). To this, Bratman's Sam replies, "'I don't think it would be best to drink. Do you think I'm stupid enough to think that, given how strong my reasons for abstaining are? I think it would be best to abstain. Still, I'm drinking'" (Bratman 1979, 156). Considering this to be the fullest version weakness of will, Bratman tries to account for it by delineating a moderate internalism.

On Bratman's view, an evaluative judgment is neither completely dissociated from nor necessarily binding on a corresponding action. Rather, the relationship is governed by a principle of rationality, which states that

> it is rational to draw a practical conclusion in favor of '*a'* from an accepted evaluative commitment in favor of '*a' unless that evaluative commitment is overridden by another evaluative commitment you accept*, or would accept if you drew all conclusions entailed by what you already accept (1979, 165-166, added emphasis mine).

In other words, according to Bratman's principle of rationality, a considered evaluative judgment results in a corresponding action unless it is 'overridden' by some other evaluative commitment. In the scenario described above, for example, Sam is able to make a full-fledged, unconditional judgment that it would be best to stop drinking, but nonetheless continues to drink because his judgment is overridden by his view that, at the moment, drinking is "quite pleasant" (1979, 156). Bratman calls this view a non-homogenous account

of the relationship between evaluative judgment and action, and he believes it provides the logical flexibility required to preserve a commitment to full-blown, considered action while also securing the logical possibility of weakness of will.

Richard Dunn takes an opposite, externalist approach to defending unconditional or 'strict' cases of weakness of will (1987). Maintaining that Davidson does not really get to the heart of the issue, he remarks,

> Davidson too is revealed as unsympathetic to the possibility of such weakness of will as it is an issue here [namely, unconditional weakness of will]. For, as I have just stressed, the concern I have with whether weakness of will is possible is specifically a concern with whether certain cases of acting against one's *unconditional* better judgment, or judgment about what is right, or some such, are possible. No doubt other putative phenomena merit being thought of it in terms of weakness of will; but none seem more central than the range of cases I have in mind; and moreover, it is surely these which, quite naturally, have provided the standard focus of discussion of whether weakness of will is possible (1987, 12).

In light of what he views as Davidson's failure, Dunn defends a complete dissociation between unconditional better judgments, on the one hand, and any volition to act on the other. Specifically, Dunn contends that there is a logical distinction between 'evaluating' and 'valuing.' The former corresponds to simply thinking about something's value, i.e., evaluating that takes place purely in the abstract (1987, 21). Only valuing implies volition. As a result, there is no logical contradiction in an agent making an unconditional, out-and-out evaluative judgment that 'Doing *x* would be terrible,' and still going on to do *x*. In other words, on Dunn's view, it is entirely possible for an agent to sincerely believe, "I ought not to steal this laptop under any circumstances, as it would be harmful and wrong," and still go on and steal the laptop in the very same moment.

Many other authors have taken up the problem of weakness of will, navigating, as Stroud describes it, "between the Scylla of an extreme internalism about evaluative judgment which would preclude the possibility of weakness of will, and the Charybdis of an extreme

externalism which would deny any privileged role to evaluative judgment in practical reasoning or rational action […]. Naturally, different theorists have plotted different courses through these shoals" (Stroud 2014, 21). Some have adopted related internalist accounts related to that of Bratman (Tenenbaum 1999, Stroud 2003, Watson 2003). Others have taken up opposing externalist approaches to unconditional weakness of will (Stocker 1979, Mele 1987, 2013). Nevertheless, although the 'next generation' of philosophers discussing weakness of will departed from Davidson with respect to the topic of unconditional judgments, it has stayed closer to his position on two other issues.

First, most philosophers agree with Davidson that weakness of will is not only logically possible, but also an actual phenomenon in everyday life (except Watson 1977). Second, philosophers have opened up the issue of whether unconditional weakness of will can be consistent with rational action (for an excellent review, see Stroud and Tappolet 2003). Most agree with Davidson that weakness of will is fundamentally irrational, although there have been prominent opponents of this view (see Audi 1990, McIntyre 1990, and Arpaly 2000).

In this way, syllogism-based models have dominated discussions of weakness of will almost continuously since Aristotle's *Nichomachean Ethics* (see also Harman *et al.* 2011). How do they stack up against Craver and Darden's criteria for evaluating mechanism schemas?

5. Assessing Syllogism-Based Accounts as Mechanism Schemas

Based on Craver and Darden's criteria for evaluating mechanism schemas, the syllogism-based models of weakness of will considered in this chapter are incomplete and incorrect.

5.1. Incompleteness

Let us start with Aristotle's account. Following Santas' interpretation (1969), Aristotle argues that a weak-willed agent is able to draw the appropriate conclusion, but she does not exercise this knowledge because she is 'overcome by passion' (1984, 1813/1147b17). From a mechanistic perspective, being 'overcome by passion' can be understood as one of the functional sub-mechanisms. But as noted in Section 3.1, it is not clear what happens when the agent is 'overcome by passion.' Aristotle's analysis thus contains at least one 'black box' that proposes a functional component but does not elucidate how it would work (Figure 2.6.).

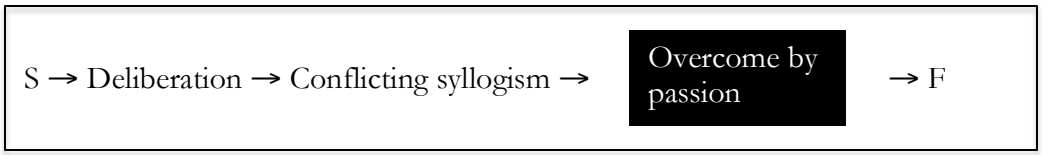

Figure 2.6. A representation of Aristotle's model of weakness of will. 'S' represents the 'starting conditions' in the mechanism schema. 'F' represents the 'finishing conditions' in the mechanism schema. Using Craver and Darden's criteria, it corresponds an incomplete mechanism schema. (adapted from Craver and Darden, 2013, 87, Figure 6.1.).

Davidson's account is characterized by a similar gap when he concludes, "the agent has a better reason for doing something else," but does the akratic act anyway (1980, 40). An uncharitable critic could argue that Davidson's position is superficial, insofar as it simply re-describes the phenomenon by saying, 'the agent had strong reasons to do *x*, but she has even stronger reasons to do *y*, so she does *y*." A more moderate reading should grant that Davidson does want to provide a detailed account of practical reasoning, and indeed does so for those cases where an agent can make an inductive inference about what might be the best thing to do. On Davidson's view, an individual does not need to make unconditional or deductive judgment about what to do; she only needs to be able to weigh all the relevant evidence available to her and make the best inference possible from it. Nevertheless, much

like Aristotle's account, it leaves a functional black box when it comes to explaining the mechanism of what actually causes the agent to pursue an expectedly less beneficial course of action (Figure 2.7.).

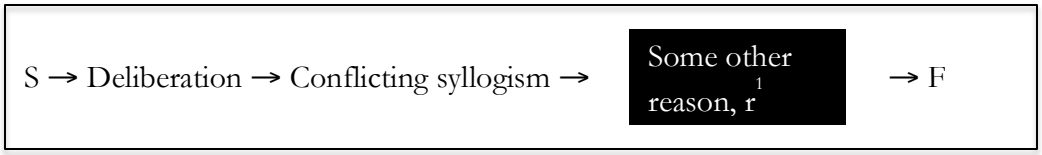

Figure 2.7. A representation of Davidson's model of weakness of will. 'S' represents the 'starting conditions' in the mechanism schema. 'F' represents the 'finishing conditions' in the mechanism schema. Using Craver and Darden's criteria, it is an incomplete mechanism schema (adapted from Craver and Darden, 2013, 87, Figure 6.1.).

Bratman and Dunn's accounts do not fare much better. Bratman uses a classic 'filler term,' of one reason "overriding" another reason to explain what happens in weakness of will, without providing any further details regarding what such an 'override' might entail. For his part, Dunn provides a detailed logical analysis for why an evaluative judgment need not entail a practical action, but does not provide any kind of a constructive account for why it turns out to be the case that most of our deliberations *do* result in appropriate actions, or what accounts for our doing something other than what our evaluative judgment would call for.

6.2. Incorrectness

Nevertheless, a mechanism schema that is on the right track can identify accurate functional sub-mechanisms and yet be unable to characterize them, thus remaining incomplete while being correct as far as it goes. In this sub-section, I want to conclude by suggesting that syllogism-based models of weakness of will are not only incomplete but also incorrect; that is, that they are directly at odds with established empirical evidence.

In a recent article entitled "Moral Reasoning," philosophers Gilbert Harman, Kelby Mason, and Walter Sinnott-Armstrong argue that although the syllogism-based model of practical reasoning has been "highly influential" throughout the history of philosophy, "as far as we know, there is no empirical evidence that people always or often form moral judgments in the way suggested" by the syllogism-base model (2011, 214, 217).[25] In a direct, empirically based critique, Harman and colleagues bring together several experiments indicating that people do not use syllogism-based reasoning to make decisions.

First, Harman and colleagues refer to a study by Cushman *et al.* (2008) suggesting that in real-life decision-making, the premises of an argument are frequently interdependent and consequently biased. On the traditional view, an agent draws on stable conceptual categories to make practical decisions; if she identifies an action that fits into an appropriate category, she acts accordingly. For example, imagine a student named Sarah debating whether or not she should cheat on an exam. Sarah may go through the following process of deliberation, with 'cheating' representing the category in question:

> I should not cheat.
> Looking at Kirstie's exam paper is cheating.
> Therefore,
> I should not look at Kirstie's exam paper.

Harman and colleagues challenge this view, however. They propose that an agent first makes moral judgment, and then this judgment determines how we categorize a specific act. On their view, Sarah's example would be more likely to play out as follows:

> I should not cheat.
> But looking at Kirstie's exam paper isn't really cheating, since I also read the book.
> Therefore,
> It is ok if I look at Kirstie's exam paper.

---

[25] Harman *et al.* call the same model the 'deductive' model, but this does not cover all uses of the model, including Davidson's inductive account in "How is Weakness of The Will Possible?" (1970). For this reason, I prefer the phrase 'syllogism-based' model and use it throughout.

Cushman *et al.*'s findings support this latter view. In the experiment, the researchers asked participants to evaluate several moral scenarios. In one paradigm, Dr. Bennett is an emergency-room physician who has an unconscious homeless man brought in. His organs are failing and he is hooked up to a life-saving respirator; the prompt states that with the respirator and some attention from Dr. Bennett, the patient may live for a week or two, but he will never regain consciousness and will certainly not live longer than two weeks.

Participants are then presented with differing descriptions of Dr. Bennett in the two conditions of the experiment. In Scenario #1, Dr. Bennett reasons, "this poor man deserves to die with dignity. He shouldn't spend his last days hooked up to such a horrible machine. The best thing to do would be to disconnect him from the machine" (2008, 283). In Scenario #2, Dr. Bennett thinks, "this bum deserves to die. He shouldn't sit here soaking up my valuable time and resources. The best thing to do would be to disconnect him from the machine" (2008, 283).

In the second half of the experiment, the participants were asked to answer questions about Dr. Bennett's actions, including (Q1) whether it is more appropriate to say that Dr. Bennett ended the homeless man's life, or that he allowed it to end, and (Q2) whether the doctor's behavior was morally wrong (Cushman *et al.* 2008, 283). Supporting Harman *et al.*s' view, Cushman *et al.* found that how people responded to Q2, namely, how they judged Dr. Bennett's behavior toward the homeless man, was directly correlated with whether they thought Dr. Bennett 'ended' the man's life, or merely 'allowed it to end.' Specifically, participants rated the morally bad doctor as having ended the patient's life significantly more frequently than they did the morally ambiguous doctor (Cushman *et al.* 2008, 284). In this way, the rules or 'categories' that we use to make judgments are thus not genuinely independent or neutral.

Harman and colleagues cite a second study by Sinnott-Armstrong, Mallon, McCoy, and Hull (2008) that further challenges the view that syllogisms are used in everyday practical reasoning. In particular, the study suggests that people do not actually employ abstract moral principles to make judgments. Although people consistently give moral reasons for their judgments or beliefs, they unconsciously use different, non-moral principles to judge scenarios.

In the study, participants were introduced to three versions of the trolley problem. In the first case, known as the 'side track case,' an agent named 'Peter' has the option of flipping a switch to ensure that a trolley moves onto a side track, thereby killing one person instead of five people being killed by the trolley otherwise (Figure 2.8).

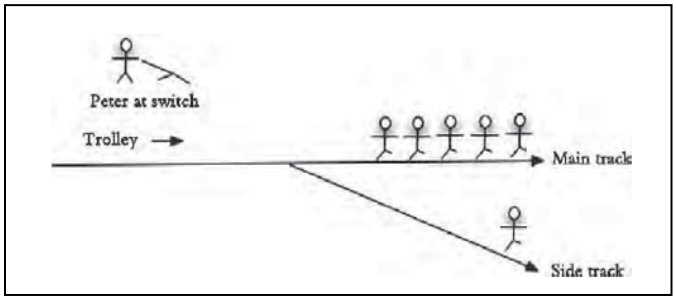

Figure 2.8. The trolley problem (side track case). In this scenario, the trolley would move onto the side track. The 5 individuals would be saved regardless of whether the single person would have been hit by the trolley or not (Adopted from Sinnott-Armstrong *et al.* 2009).

In the second case, known as the 'loop track case,' Peter has the option of flipping a switch to redirect the trolley onto a side loop, where it will kill one person but also stop, thereby saving the lives of five other people (Figure 2.9). In the third and final case, Peter has the option of flipping a switch to redirect the trolley onto a side loop, before flipping a second switch to guide the trolley onto a side track. As in the previous two cases, Peter's actions will result in the death of one individual, but prevent the death of five others (Figure 10). In each case, the participants were asked about Peter's action, namely, whether (Q1) Peter 'killed' the solitary individual and (Q2) it was 'morally wrong' for Peter to act as he did.

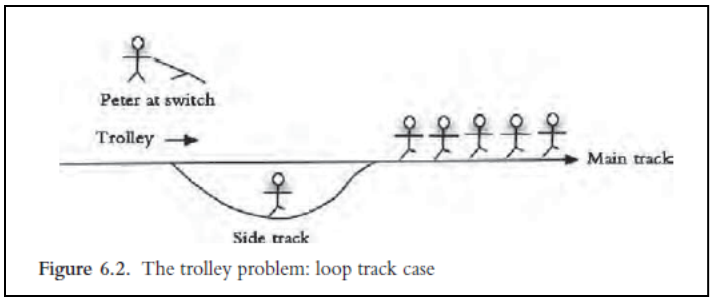

Figure 6.2. The trolley problem: loop track case

Figure 2.9. The trolley problem (loop track case). In this scenario, the trolley would move onto the loop track. The 5 individuals would only be saved if the single individual were to be hit by the trolley first (Adopted from Sinnott-Armstrong *et al.* 2009).

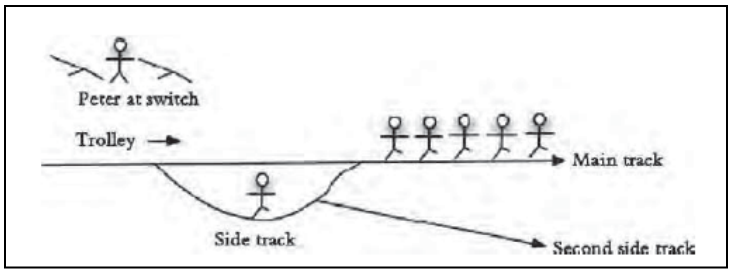

Figure 2.10. The trolley problem (combination track case) In this scenario, the trolley would first move onto the loop track; if the second witch were changed, the trolley would move onto the second side track. The 5 individuals would only be saved if single person were to be hit by the trolley *and* the trolley is moved onto the second side track (Sinnott-Armstrong *et al.* 2009).

The results of the study were striking. When asked to justify their views, participants who thought that Peter's actions were wrong typically responded by suggesting, "Peter's actions were morally wrong because it is wrong to kill someone" (2008). But the results indicated that participants' judgments about the moral status of Peter's actions were far more significantly influenced by the entirely non-moral factors of temporal order and an intentional to kill (as opposed to killing as a side effect) than they were by any moral considerations.

On one hand, participants were far more likely to answer that Peter had killed the individual if the death of the individual occurred *before* the other five were saved than if it did not. In other words, participants felt that Peter had killed the individual in the loop track and

combination track cases, where the single individual dies before the safety of the other 5 is assured, but not in the side track case, where the group is saved the instant the trolley switches tracks (and the individual only dies at some point after that). In other words, in the first two cases, the solitary individual has to die before the group of five is positively safe. By contrast, in the side track case, the 'lifesaving' switching of the track occurs before the individual is run over. This latter scenario was closely correlated with participants responding that Peter had not, in fact, killed the individual. This finding led Sinnott-Armstrong *et al.* conclude that the temporal order of events, and not moral principles, played a defining role in whether participants deemed the events to involve killing or not.

On the other hand, the researchers found that participants only judged Peter's actions to be morally wrong if he intended to kill the individual, that is, if he needed the body of the single individual to slow the movement of the trolley. Correspondingly, participants were far more likely to respond that Peter's actions were morally wrong in the loop track case than they were in the side track and combination cases, where the trolley could be stopped without needing the body to slow it down. In these latter two cases, the extra track would have been sufficient to stop the trolley, but the additional individual was simply in the way.

Together, these two pieces of evidence contradict the majority of participants' view that they judge moral responsibility based on the fact that killing is wrong. If this had been the case, they would have singled out the same cases as "killing" and to be "wrong." But that is not what the researchers found. The participants thought that Peter had killed the individual on the loop track and combination cases; they only thought his actions were morally wrong on the loop track case. This suggests that people use entirely different factors to make moral judgments than those they consciously invoke. This finding provides Harman

et al. (2011) further proof that syllogism-based reasoning is not actually used by individuals in every day formulation of judgments and decision-making.

6.Conclusion

Since practical reasoning provides the framework for weakness of will, it is likewise unlikely that syllogism-based reasoning is used in cases resulting in weakness of will. The evidence discussed in the preceding section supports the view that the syllogism-based model is not only explanatorily incomplete but also incorrect. With this added empirical perspective, it can be argued that – no matter how prevalent and influential – syllogism-based models of weakness of will constitute an incorrect schema of the mechanism underlying weakness of will.

I suggest that syllogism-based models cannot explain how an agent chooses between two competing alternatives because they lack any sub-mechanism whereby certain propositions can be identified as more important or pursuit-worthy than others. Without a unit of value or some other means to 'weigh' them, nothing can 'tip the scales' in a contest between two contradictory or at least somehow competing propositions. As a result, an agent's consideration of competing alternatives results in a logical stalemate. In those rare cases that the stalemate is resolved, as it is in some accounts of weakness of will, it is as the upshot of a mysterious mechanism as, for example, in Davidson's suggestion that "there simply is a reason" for an agent's moment of weakness of will (1980, 40).

CHAPTER 3

THE ALTERNATIVE POSITION: VALUATION-BASED MODELS OF WEAKNESS OF WILL

1. Introduction

In Chapter 2, "The Standard Theory: Syllogism-Based Explanations of Weakness of Will," I argued that although syllogism-based models of weakness of will have historically been influential, they provides only a limited account of how decision-making and weakness of will take place. Examining how philosophers from Aristotle to Michael Bratman have deployed different versions of the syllogism-based model, I referred to Craver and Darden's criteria for evaluating mechanism schemas to show that syllogism-based accounts are not only incomplete but also incorrect. I further suggested that their accounts could not explain how an agent chooses between two competing alternatives, or how an agent could first choose one course of action but then go on to pursue another, because the syllogistic theory does not account for how certain propositions can be identified as more important or pursuit-worthy than others.

Historically, a second prominent model for explaining weakness of will is in terms of conflict between different faculties, and particularly between reason and the affects (or emotions). On this view, a moment of weakness of will occurs when an agent possesses rational knowledge regarding what it would be best to do, but is overcome by a competing, irrational faculty that prevents her from doing so. One important representation of this account is Plato's discussion of the tripartite soul in Book IV of the *Republic*, where he suggests that the rational, spirited, and appetitive parts may occasionally conflict with one another (Cooper 1997/444b, translated by G.M.A. Grube, revised by C.D.C. Reeve). Other philosophers who have taken up versions of the conflict-based approach to weakness of will

include Hume (Fleming 2010) and Kant (Hill Jr. 2008, 2012) and, more recently, Bennett (1974) and Tappolet (2003).

In this chapter, I highlight key components of conflict-based models of explaining weakness of will, but I do it from the perspective of a new interpretive model, based on the mechanisms of 'valuation.' The concept of valuation refers to the processes whereby we come to value and seek out what benefits us as living organisms. Valuation-based models of weakness of will typically stipulate:

Valence	That agents attribute values to internal and external objects and events,
Activation	That positively valuated objects and events elicit approach responses, while negatively valuated objects and events elicit withdrawal responses and, finally,
Error	That error is the product of agents evaluating an alternative as *apparently* more valuable than it actually is.

In focusing on values and valuation, I depart from standard syllogism- or conflict-based interpretations of weakness of will. At least three reasons motivate this interpretive shift, presented in order of increasing importance. First, a valuation-based model allows me to discuss several historical positions that are typically less represented in the mainstream debates; second, these somewhat less familiar positions contain important theoretical resources for dealing with the problem of weakness of will; and finally, a valuation based model will provide a useful heuristic for the computation-based accounts developed in Part II of the dissertation.

The remainder of the current chapter is devoted to a detailed analysis of four historical, valuation-based models of weakness of will. In particular, I examine the positions of Plato's Socrates (in both the *Protagoras* and *Republic*), Spinoza, and R.M. Hare. I argue that,

although these authors clearly set out from different philosophical points of departure, they each endorse the principles of valence, activation, and error.

In Section 2, I begin by rejecting the commonly held view that Plato's Socrates denies that the phenomenon of weakness of will exists, and argue that he merely criticizes the multitude's characterization of being 'overcome by pleasure. Turning to the latter portion of the *Protagoras*, I then go on to reconstruct an alternative, naturalistic explanation of weakness of will based on Socrates' discussion of valuation, activation, and error (or 'ignorance').

Many scholars have argued that Plato departs from his commitment to a unified soul in the *Protagoras* to a more mature account in the *Republic*. In Section 3, I discuss Plato's discussion of the tripartite soul in the *Republic* and consider the implications of this position for the problem of weakness of will. I pay particular attention to Part IX of the *Republic*, where Plato's Socrates argues that each part of the soul has its own pleasures and desires. On the basis of my analysis, I suggest that Plato presents an early, multi-system account of practical reasoning and weakness of will.

In Section 4, I show how Spinoza outlines and defends a causal psychological theory of weakness of will or, as he calls it, "the causes of human weakness and inconstancy" (IV P18 *Scholium*). I argue that for Spinoza, the issue takes on an added significance, because he is the first philosopher to explain weakness of will without resorting to a concept of free will; rather, he can provide an explanation of the phenomenon not 'in spite of' but precisely *because* of his rejection of a notion of free will. In doing so, Spinoza not only adopts weakness of will as a test case of his psychological theory, but also uses it to magnify and explain his general understanding of the relative relationship between knowledge and the emotions.

In Section 5, I argue that Hare's explanation of weakness of will is consistent with two of the three central elements of an evaluative position, namely, the principles of valuation and activation. I deny, just as I did in the case of Plato's Socrates, that Hare rejects the possibility of weakness of will in light of his prescriptivism. Rather, I propose that Hare is careful to maintain theoretical flexibility within his prescriptivist position, making space for an 'ought but can't' position in which the agent knows what she ought to do, but is either physically or psychologically incapable of doing it.

Finally, in Section 6, I use Craver and Darden's criteria to argue that the valuation-based models discussed underspecify key functional sub-mechanisms, but are consistent with current empirical evidence. I suggest that they are at a 'how-possibly' stage of correctness.

But let us start by turning to the oldest known discussions of weakness of will, namely, in Plato's *Protagoras*.

2. Weakness of Will in Plato's *Protagoras*

Plato's dialogue the *Protagoras* is devoted to an analysis of virtue, and in it the figure of Socrates endeavors to show that all of the virtues are inseparable and coincide with knowledge (or wisdom). His interlocutor in the dialogue, Protagoras, wants to maintain that courage is different from the other virtues and can be possessed separately from them. It is in this context that Socrates sets out to defend an intermediate thesis that brings him to the discussion of weakness of will. Rejecting the view that "while knowledge is often present in a man, what rules him is not knowledge but rather anything else – sometimes anger, sometimes pleasure, sometimes pain, at other times love, often fear," Socrates sets out to identify control of oneself with wisdom, and lack of control with ignorance (352b-c, 358b, translated by Stanley Lombardo and Karen Bell).

Following from within this frame, Socrates's examination of weakness of will can be divided into two main branches, each of which contains a central argument supported by several sub-arguments:

353e - 355a	355b-358e
'Critical branch'	'Constructive branch'
To prove multitude's view of 'being overcome by pleasure' as absurd	To explain weakness of will as the product of ignorance

In what I will call the 'critical branch' of the argument, presented between 353e and 355a of the *Protagoras*, Socrates discusses the multitude's view that weakness of will is the result of 'being overcome by pleasure.' Socrates criticizes the multitude's reasoning as being self-contradictory and ultimately as having necessarily absurd implications. Socrates does not, however, as it is often charged, deny the existence of weakness of will altogether. Returning to the terminology I borrowed from Christopher Shields in Chapter 1, Socrates recognizes the relatively regular occurrence of 'implementation failure' in everyday human life; he only rejects the multitude's explanation of what causes this phenomenon.

In the second, 'constructive branch' of his argument, presented between 355b and 358e, Socrates sets out his alternative explanation of weakness of will, which famously leads him to conclude that weakness of will is the product of ignorance. A brief analysis of this second branch shows Socrates's analysis to be in line with the main features of valuation-based models of weakness of will. I reconstruct each of these two branches in turn.

2.1. The Critical Argument

Socrates opens his critical argument by establishing the hedonistic thesis that pleasure is good and pain is bad.[26] He then considers the multitude's statement regarding self-control, which suggests, "frequently a man, knowing the bad to be bad, nevertheless does that very thing, when he is able not to do it, having been driven and overwhelmed by pleasure" (355b). Socrates shows that this statement is absurd in two ways. First, he draws on the hedonistic thesis to substitute every instance of 'pleasure' with 'good,' so that the multitude's statement now reads, "a man knowing the bad to be bad, nevertheless does that very thing," because he is overcome by the good (355c). But this proves the statement is, according to Socrates, "ridiculous," since the good and the bad are objectively quantifiable – "one is greater and one smaller, or more and less" (355e) – and as such, the multitude's statement would mean that a man is 'overcome' insofar as he purposely sets out to get more bad things for the sake of fewer good ones.

Second, Socrates reverses the exchange and substitutes every instance of 'good' with 'pleasure.' The multitude's statement now describes a man who "does what before we called 'bad' things and now shall call 'painful' ones, knowing they are painful things, but being overcome by pleasant things, although it is clear that they do not outweigh them" (356a). Once again, Socrates appeals to the objectively quantifiable nature of pleasures and pains to argue that the multitude's statement would mean that a man is 'overcome' insofar as he chooses more and more intensely painful things for the sake of fewer and less intensely pleasurable things. In this way, Socrates uses both substitutions to prove that the multitude's position regarding the power of pleasure results in an absurd conclusion.

[26] This is a problem. See Section 2.1.

Again, however, it is worth emphasizing that Socrates does not deny the existence of weakness of will, i.e., implementation failure. He rejects the view that another faculty can overcome knowledge, but acknowledges that the experience of not doing what is best occurs on a regular basis in everyday life. He explicitly says to Protagoras,

> Come with me, then, and let's try and persuade people and to teach them what is this experience which they call being overcome by pleasure, because of which they fail to do the best thing when they know what it is. For perhaps if we told them that what they were saying isn't true [i.e., that this experience is caused by being overcome by pleasure], but is demonstrably false, they would ask us: 'Protagoras and Socrates, if this is not the experience of being overcome by pleasure, but something other than that, what do you two say it is? Tell us" (353a).

It is not that the experience doesn't really exist – it is that the multitude's explanation of it is incorrect or 'false.' And indeed, as promised in the passage, Socrates does go on to "tell us" what this experience is really caused by.

2.1. The Constructive Argument

The constructive branch of Socrates' analysis is more straightforward than its negative counterpart. Here, Socrates presents his positive view that weakness of will is, in fact, caused by a lack of knowledge about what is genuinely pleasurable and what is genuinely painful. He begins by reiterating his view that pleasures and pain are objectively quantifiable, and explains that they "are not different in any other way than by pleasure and pain, for there is no other way that they could differ" (356b). He then offers the analogy of weighing. Valuation in decision-making is like weighing pros and cons on a scale: "you put the pleasures together and the pains together, both the near and the remote, on the balance scale, and then say which of the two is more" (356b). Socrates then further specifies that

human beings naturally choose what is more pleasant and less painful, and seek out what is pleasant and avoid what is painful (356b-c).

We can identify these as two distinct principles representing the basic laws in of his theory of motivation. Socrates phrases them as follows:

> (L1) "If you weigh pleasant things against pleasant, the greater and the more must always be taken; if painful things against painful, the fewer and the smaller," and
>
> (L2) "If you weigh pleasant things against painful, and the painful is exceeded by the pleasant – whether the near by the remote or the remote by the near – you have to perform that action in which the pleasant prevails; on the other hand, if the pleasant is exceeded by the painful, you have to refrain from doing that" (356b-c).

Together, he draws on both of these principles to reason that those who choose less pleasure and/or more pain must do so out of ignorance – that is, they must be ignorant of what is genuinely more pleasurable and less painful, or again, what is good and what is bad (357d-e). To advance any other explanation, including the suggestion that the weak-willed individual is 'overcome by pleasure,' would entail a violation of human nature, and so would be absurd or, as he puts it, "ridiculous" (355d).

2.3. The Problem of Hedonism

For the historian of philosophy, *Protagoras* 351-359 nevertheless requires clarifying the exact status of the thorny hedonism thesis. After all, how can Plato advance a *hedonistic* account?

Several prominent scholars have struggled to resolve this issue. A number have argued that the hedonistic thesis likely represents the real views of the historical Socrates (Adam, 1893), or even Plato's own views at an early stage of his philosophical development (Hackforth, 1928, Vlastos, 1956, to a certain extent, Vlastos, 1969, Klosko, 1980, Reeve 1992, Irwin 1995). Others have suggested that the whole discussion in the Protagoras is

intended as an ad hominem attack on the Sophists, and that the hedonistic position serves as nothing more than a throwaway premise on the way to proving that virtue is knowledge (Grube, 1933; Sullivan, 1961). A third group has denied that Socrates' views amount to a hedonistic account in any meaningful sense (Goodell, 1921).

Following Vlastos (1969, 76), one can suggest that Socrates himself subscribes to a moderate version of the claim, namely claim (A):

(A) all pleasure is good and all pain is bad,

while the 'multitude' subscribes to a much stronger version of the hedonistic thesis, according to which:

(B) all pleasure is good and all pain is bad, and all good is pleasure and all evil is pain, or in other words, the good is nothing other than pleasure, and the bad is nothing other than pain (355a).

What implications does this distinction have for Socrates' analysis of weakness of will? Is the stronger version of these claims *essential* to the Platonic conception weakness of will, or is it only misleadingly interwoven with it? Here, it is helpful to keep the two branches of Socrates' argument apart.

I contend that the negative branch of Socrates' argument, i.e., that weakness of will is not caused by 'being overcome' with pleasure, is consistent with claim (A), but also further *requires* its stronger counterpart, claim (B). Specifically, Socrates uses (L1) and (L2), together with claim (B), to show that the phenomenon of being 'overcome by pleasure' would entail a violation of human nature, and so must be absurd, or "ridiculous" (355d). Socrates' critical sub-argument rejecting the view of the multitude may be reconstructed in the following way.

M a man does evil, knowing that he does evil, because he is 'overcome by pleasure' (the multitude's thesis)

P1 (B) all pleasure is good and all pain is bad, and all good is pleasure and all evil is pain, or in other words, the good is nothing other than pleasure, and the bad is nothing other than pain

P2 (L1) "if you weigh pleasant things against pleasant, the greater and the more must always be taken; if painful things against painful, the fewer and the smaller"

P3 (L2) "if you weigh pleasant things against painful, and the painful is exceeded by the pleasant – whether the near by the remote or the remote by the near – you have to perform that action in which the pleasant prevails; on the other hand, if the pleasant is exceeded by the painful, you have to refrain from doing that"

P4 Good and bad must be compared with one another "either as greater and smaller, or [as] more and fewer"

1. By (P1), 'good' may be substituted with 'pleasure,' and 'bad' may be substituted by 'pain'
2. By (P1) and (1), (A) can be said to read, 'a man does evil, knowing that it is evil, because he is overcome by good'
3. By (P4) and (2), (A) can be said to read, 'a man does evil, knowing that it is evil, because he chooses the greater evil in exchange for the lesser good?'
4. By (P1), 'pleasure' may be substituted with 'good,' and 'pain' may be substituted by 'bad'
5. By (2) and (4), (A) can be said to read, 'a man does what is painful, knowing that it is painful, because he chooses what is more painful in exchange for what is less pleasant'
6. But (5) contradicts (P3)/(L2)

Therefore, ~M

In this way, Socrates contradicts the multitude's statement that weakness of will consists in being 'overcome by pleasure.'

Socrates does not need claim (B) to prove his positive conclusion that 'weakness of will is ignorance,' but he does rely on it to demonstrate the further principle that "no one goes willingly toward the bad or what he believes to be bad" (358d). This second sub-argument may be reconstructed as follows:

P1 (B) all pleasure is good and all pain is bad, and all good is pleasure and all evil is pain, or in other words, the good is nothing other than pleasure, and the bad is nothing other than pain

P2 (L1) "if you weigh pleasant things against pleasant, the greater and the more must always be taken; if painful things against painful, the fewer and the smaller"

P3 (L2) "if you weigh pleasant things against painful, and the painful is exceeded by the pleasant – whether the near by the remote or the remote by the near – you have to perform that action in which the pleasant prevails; on the other hand, if the pleasant is exceeded by the painful, you have to refrain from doing that"

P4 ($L2^1$) Knowing means one must pursue what is less painful and more pleasant, or again, no one knowingly pursues what is more painful for what is less pleasant

1. By ($L2^1$), No one knowingly pursues what is more painful for what is less pleasant
2. By (P1), 'pleasant' may be substituted with 'good,' and 'painful' may be substituted by 'bad'
3. By (1) and (2), "no one goes willingly toward the bad or what he believes to be bad"

Thus, Socrates does not require the stronger version hedonistic thesis (claim (B)) to formulate his basic account of weakness of will – but he does use it to make a statement about human beings' tendencies to commit bad actions. Interestingly, Socrates also uses his positive argument to infer a third law, namely that,

(L3) "To prefer evil to good is not in human nature; and when a man is compelled to choose one of two evils, no one will choose the greater when he may have the less" (358d).

In this way, Plato's Socrates relies on a pair of hedonistic theses as the basis for a valuation-based account of weakness of will.

Fortunately, two further factors help moderate the problematic possibility of a Platonic endorsement of hedonism. First, Plato appears to depart from the views presented in the *Protagoras* to outline a more mature account in the *Republic*. In particular, he discusses the nature of the tripartite soul and presents an alternate account of weakness of will that does not depend on a commitment to hedonism. I discuss both of these developments in Section 3.

Second, from an ahistorical, naturalist perspective, we do not need to be quite as worried about Plato's possible endorsement of hedonism. Rather, what should be striking

here is how Socrates' principles (L1-L3) anticipate a valuation-based (or what is also known as an 'appraisal oriented' (Arnold 1960, Lazarus 1991, Fridja 1986, Scherer 2006) theory of valence, attraction, and aversion. I use Craver and Darden's criteria for evaluating mechanism schemas to discuss the relative naturalism of Plato's views in Section 5 below.

3. The Tripartite Soul and Weakness of Will in the *Republic*

In Part IV of the *Republic,* Plato returns to the discussion of appetite and the good that had previously been taken up in the *Protagoras*. In this later dialogue, however, Plato's Socrates appears to endorse a very different position. He remarks, "Let no one catch us unprepared or disturb us by claiming that no one has an appetite for drink, but rather good drink, nor food but good food, on the grounds that everyone after all has appetite for good things" (1984/438a). With these words, Plato's seems to reject his earlier thesis that our appetites pursue 'the good,' and not specific objects such as drinks or food (Irwin 2005, 206).[27] Individuals do not pursue pleasure or what is good for its own sake; rather, they desire a wide variety of things, and some of these also happen to be pleasurable and good.

Plato thus appears to reject the hedonistic principles endorsed in the *Protagoras*. So what becomes of his discussion of weakness of will? I suggest that in the *Republic*, Plato presents an alternate, valued-based account of weakness of will based on his conception of the tripartite soul. In particular, the nature of the tripartite soul allows him to argue that different parts of the soul value different things, bringing them into a conflict that causes weakness of will.

---

[27] Many scholars interpret this point as the "the precise place at which Plato departs from the Socrates of the earlier dialogues regarding the psychology of choice, that is, regarding how a person chooses among the various alternatives that are open to him" (Weiss 2007, 87; see Watson 1977, Cooper 1984, Brickhouse and Smith 1994, Miller 1999, Reeve 1992, Irwin 2005. Weiss is critical of this view, however, as is Shields 2007).

Famously, the tripartite soul is made up of rational, spirited, and appetitive parts. The purpose of the rational part is to pursue truth and to rule the individual; the spirited part follows the rule of the rational part, but also pursues victory and honor; and finally, the appetitive part seeks bodily pleasures, and is also called 'money-loving' (441e-442). Plato uses the tripartite soul to define justice for the individual just as he does for the city, suggesting that "one who is just does not allow any part of himself to do the work of another part or allow the various classes within him to meddle with each other. He regulates well what is really his own and rules himself" (443c-d). At the same time, the tripartite soul enables Plato to provide a more straightforward account of weakness of will than he was able to in the *Protagoras*.

In Book IX of the *Republic*, Plato's Socrates suggests that there are not only different parts of the soul, but that each part of the soul has different desires or values. He observes, "it seems to me that there are three pleasures corresponding to the three parts of the soul, on peculiar to each part, and similarly with desires and kinds of rule" (580d). He then goes on to specify that the appetitive part seeks food, drink, and sex. The spirited part pursues control, victory, and reputation. For its part, the rational component of the soul pursues truth (580e-581c). In turn, the different desires of each part can cause them to come into conflict with one another.

Describing a kind of "civil war" between the different parts, Socrates explains that it is possible for the lower, i.e., appetitive part of the soul to disobey the natural rulings of the rational part, and thereby cause an individual to pursue a wide variety of vices (444b). It is equally possible for the appetitive part to overpower the spirited part. To illustrate this, Socrates uses the example of Leontius both wanting and not wanting to look at the bodies of individuals who had recently been executed. He describes how "Leontius saw some corpses

lying at the executioner's feet. He had an appetite to look at them but at the same time he was disgusted and turned away. For a time he struggled with himself and covered his face, but finally, overpowered by the appetite, he pushed his eyes wide open and rushed towards the corpses, saying, 'Look for yourselves, you evil wretches, take your fill of the beautiful sight!'" (339e-440a). In this way, the appetitive part overwhelms the spirited part, which then rushes up as an experience of anger.

Plato's discussion of the tripartite soul thus allows him to present an alternate account of weakness of will that does not depend on a commitment to hedonism, but instead broadly commits him to a value-based account.

4. Spinoza's Theory of Weakness of Will

In this section, I will show how Spinoza outlines, tests, and defends his own innovative, causal psychological theory by explaining weakness of will or, as he calls it, "the causes of human weakness and inconstancy" (IV P18 *Scholium*).[28] Specifically, I show how Spinoza can explain weakness of will without relying on a concept of free will, the major resource in medieval discussions of weakness of will.[29] He presents a theory based on what he understands to be the fundamental relationship between the relative motivational force of knowledge and the emotions.

---

[28] *Spinoza: Complete Works,* Shirley (Trans.), 2006.

[29] It is worth noting that Spinoza's rejection of an independent faculty of the will brings him very close to many contemporary cognitive scientists' position on this issue (Libet 1985, Wegner 2002, Wegner 2003, Custers and Aarts 2010). This has led several philosophers to see Spinoza as a pioneer of modern scientific psychology. For example, Heidi Morrison Ravven argues, "recent evidence suggests that Spinoza may have gotten it right" (2003). And despite the predictable difficulties of drawing on a relatively neglected and challenging model, William Meehan, a clinician, argues that "like anyone else, Spinoza becomes more accessible with familiarity, and the value of acquiring that familiarity is evidenced in the remarkable extent to which his insights and observations anticipated the findings of contemporary neuroscience and those of a variety of psychologists and philosophers of science. To understand Spinoza, I argue, is to understand, focus and enrich a paradigm shift that has already begun" (2009). Although this is a tempting line of reasoning, I explain why I think it is an oversimplification in Section 6 below.

Spinoza rejects the voluntarist theories of weakness of will brought forward by the later medieval philosophers. Yet he also turns away from the classic 'belief and desire' model that had played an important role in discussions of weakness of will since Plato. Instead, Spinoza draws on his conceptions of knowledge and the emotions to articulate a deterministic, valuation-based theory of weakness of will.

4.1. Rejecting the faculty of the will

Spinoza frames his discussion of weakness of will with an extended critique of the later medieval and Cartesian distinction between the intellect and the will. He denies that there is any distinction between the intellect and the will (II P 49 corollary), and he further rejects any interpretations of weakness of will which would suggest that it is the product of a conflict between them. He explicitly challenges these interpretations of weakness of will in his Preface to Part III, and writes that most "assign the cause of human weakness and frailty not to the power of Nature in general, but to some defect in human nature, which they therefore bemoan, ridicule, despise, or as is most frequently the case, abuse," (and both here and in Part IV, he uses the terms "*impotentiae et inconstantiae*" to refer to weakness of will). Since Spinoza's discussions of the will and free will have been the subject of several careful and systematic treatments, including those of Cottingham (1988) and Lloyd (1990), the latter of which I adopt here, I will turn directly to the implications of Spinoza's rejection of the will as separate from the intellect for his theory of weakness of will, which relies on an understanding of the relative forces of knowledge and the emotions.

4.2. The Big Picture And Some Basic Definitions

In general terms, Spinoza argues that our limited situation in nature means that the force of our externally-caused emotions is powerful enough to overcome our self-caused true knowledge and the related active emotions, and he explains how and why this is the case in Part IV of the *Ethics*. To do so, Spinoza draws on the concepts of imaginings and adequate ideas, passivity and activity, and the respective passive and active emotions to provide a core causal mechanism for the phenomenon of weakness of will.[30] Before looking at the details of his account, it will be useful to discuss each of these concepts in turn.

4.2.1. Adequate and Inadequate Ideas

Spinoza distinguishes between four modes of perception or knowing. The first type of knowing is obtained from hearsay. It is knowledge gained through communication with others (*TdIE* 20). The second mode of perception is obtained through passive, uncritical experience. This kind of knowledge is "not determined by the intellect, but is so called because it chances thus to occur" (*TdIE* 19). Spinoza acknowledges that "almost everything that is of practical use in life" is learned through casual experience, but evidently and nevertheless maintains that it is not a secure means for gaining knowledge: "[besides] its considerable uncertainty and indefiniteness," he explains, "no one will in this way perceive anything in natural things except their accidents" (*TdIE* 27). The third type of knowledge consists in knowing the essence of a thing, but inadequately. This kind of perception occurs "when we infer a cause from some effect or when an inference is made from some universal

[30] *Treatise on the Emendation of the Intellect (TdIE)* 33. In a letter to Ehrenfried Walter von Tschirnhaus, Spinoza makes a careful distinction between what he calls 'true' and 'adequate' ideas, explaining, "I recognize no difference but this, that the word 'true' has regard only to the agreement of the idea with its object (ideatum), whereas the word 'adequate' has regard to the nature of the idea in itself. Thus there is no real difference between a true and an adequate idea except for this extrinsic relation" (Letter 60). In other words, the designation of an idea as a true idea refers to the degree of correspondence between an object and its idea. By contrast, an adequate idea refers to the idea's own internal standard of certainty.

which is always accompanied by some property" (*TdIE* 19). For example, we exercise this kind of reasoning when we understand the principles of perspective and subsequently infer that the sun is distant from and larger in size than the earth.

The fourth, and for Spinoza most important mode of knowledge, is the perception of a thing "through its essence alone, or through its proximate cause" (*TdIE* 19). For instance, to know that the sum of two and three is five is an example of this kind of knowledge, and teach us "what it is to know something" (*TdIE* 21). Although Spinoza acknowledges that we only know a few things in this way, only this fourth mode of knowledge allows us to "comprehend the essence of the thing, and is therefore without danger of error" (*TdIE* 29). The first three kinds of knowledge involve inadequate ideas. The fourth kind of knowledge involves adequate ideas.

4.2.2. Passivity and Activity

Spinoza's theory of the emotions is based on the concept that each individual possesses a certain 'power of activity,' or force of existence (*vis existendi*). Spinoza refers to this power as the thing's essence or 'conatus,' whereby "each thing, in so far as it is in itself, endeavors to persist in its own being" (*E*III, P4, Preface to Part IV). He explains, "the human body can be affected in many ways by which its power of activity is increased or diminished; and also in many other ways which neither increase nor diminish its power of activity" (*E*III, Post 1). In this way, every individual's power of activity fluctuates over the course of both her everyday experiences and her life, according to her interactions with the surrounding environment. By the principle of parallelism, these fluctuations in the body's power of activity are paralleled in the mind (*E*III, D3). As such, when an external body affects an individual's own body, he or she is conscious of the idea of this affection in his or her mind;

and when this affection of the body increases or decreases his or her *vis existendi*, it is mirrored by an *idea* of this increase or decrease, which, descriptively, is experienced as a certain kind of emotion, namely, as an emotion of pleasure or pain. Spinoza explains, "we see then that the mind can undergo considerable changes, and can pass now to a state of greater perfection, now to one of less perfection, and it is these passive transitions (passiones) that explicate for us the emotions of Pleasure (leatitia) and Pain (tristitia)" (*E*III, P11, Sch.). 'Pleasure' consists in the "passive transition of the mind to a state of greater perfection," while 'pain' corresponds to the "passive transition of the mind to a state of less perfection" (*E*III, P11, Sch.). When the mind is conscious of its conatus, i.e., striving or desire, it experiences what Spinoza calls 'will,' and he identifies these three emotions – pleasure, pain and desire – as the three primary emotions of human experience.

4.2.3. Looking Ahead

Spinoza's argument centers on Propositions IV P1, 7, 8, 14, and 15, where he demonstrates why the emotions related to inadequate ideas necessarily have stronger motivational force than knowledge does. He develops his account in three stages: first, by discussing the relative power of true and false concepts in our imagination; second, by determining the relative motivational force of different emotions; and finally, combining the above, by demonstrating the uneven balance of power between knowledge and the emotions.

4.3. The Relative Power Of Inadequate And Adequate Ideas

Spinoza discusses the relative powers of inadequate and adequate knowledge in Part IV, Proposition 1 of the *Ethics*. He states, "nothing positive contained in a false idea is annulled by the presence of what is true, insofar as it is true." This means that if we have inadequate

knowledge of something, it is not just a matter of gaining adequate knowledge of it to remove the confused experience. On the contrary, in this proposition, Spinoza demonstrates that inadequate knowledge can frequently impact and motivate us more strongly than adequate knowledge can.

To illustrate this relationship, Spinoza describes the experience of looking at the sun. At first glance, the sun appears to be relatively close to the earth, and until we know otherwise, for example, as young children, we tend to think of the sun as being quite small. But the question is, 'What happens when we come to know the truth about the distance and size of the sun?' Interestingly, Spinoza argues that even when we know how far the sun is from the earth, we "shall nevertheless see it as being close to us" (IV P1). This is because our new-found knowledge replaces the *factual* error, but it cannot remove what Spinoza calls the 'imagining' [*imaginatio*], i.e., the sensory effect that the sun has on our body and, by the principle of parallelism, on our mind. According to II P17, these confused sensory imaginings can only be removed when they are replaced by other imaginings. So, in the balance of power, adequate knowledge fares less well than the concept superficially [implies], and adequate knowledge can and frequently is overpowered by inadequate knowledge or, in other words, by confused ideas.[31]

4.4. The Relative Motivational Force Of Different Emotions

In Propositions 2-7, Spinoza then outlines principles regarding the relative power of different emotions. He argues that the power of an emotion is determined not just by the power of the conatus of the individual experiencing the emotion, but also primarily by the power of the external causes that produce that emotion in the individual. Since human

[31] Spinoza anchors this analysis primarily in his analysis of knowledge in Part II of the *Ethics*, and the analogy of the sun goes as far back as the *Treatise on the Emendation of the Intellect (TdIE).*

beings are infinitely limited in proportion to the forces of nature (Part IV, Axiom 1), we are very rarely the adequate causes of our actions, and the vast majority of our emotions are produced at least in part by external forces and frequently overpower us. In addition, Spinoza further demonstrates that an emotion can only be checked or destroyed by an opposite and more powerful emotion. He proves this by turning back to the principles of relative *physical* forces in Part II. Accordingly, just as an affection of the body can only be checked or destroyed by an opposite and stronger corporeal cause, by the principle of parallelism, an emotion can only be checked or destroyed by an opposite and stronger emotion.

But Spinoza wants to show the interplay of power between knowledge and the emotions. That is how Spinoza will conceive of human inconstancy or weakness of will, and he establishes how this mechanism works in Propositions 8, 14, and 15.

4.5. The Asymmetrical Balance Of Power Between Knowledge And The Emotions

In a moment of weakness of will, the knowledge involved consists in true knowledge regarding what is good and bad. Since Spinoza defines 'good' as something which we know to be useful to us, and 'bad' as something which we know to be the opposite, knowledge of good and bad simply corresponds to the idea or the consciousness that we have of a certain pleasure or pain (IV D1-2, P8). By extension, *true* knowledge of good and evil corresponds to adequate ideas about what is useful and what is harmful to us. It is this knowledge which is overcome in a moment of weakness of will because, despite being adequate, even this true knowledge can be overcome by an emotion. This is due not to the relative adequacy of the knowledge and emotions involved, but rather due to their relative motivational force or

power. Specifically, Spinoza's understanding of power necessitates that adequate knowledge can only be as powerful as the limited, finite essence that produces it.

Spinoza's argument turns on the fact that, as knowledge of good and evil, this knowledge nevertheless remains "nothing other than the" emotion of pleasure and pain, insofar as we are conscious of it (IV P8). Specifically, in Proposition IV P8, he explains that the relevant knowledge or idea is "united to the emotion [of pleasure or pain] in the same way that the mind is united to the body," that is, that it is different "only in conception." As a result, according to III *Def. 1*, as an emotion, this knowledge also produces a desire proportional to its force in the mind. The problem is that, since the individual is the sole and adequate cause of her true knowledge, only her specific, limited, and individual essence produces that desire and, as a result, defines, i.e. limits its relative power (IV P5). By contrast, the bulk of our emotions consist in passive experiences, which are generated by powerful external causes, and thus considerably surpass us in force (IV P3). As a consequence, our passive emotions can and frequently do overpower even our true knowledge (IV P14). And so, even if an individual possesses true knowledge regarding what would be a beneficial or detrimental course of action to pursue, her passive emotions may very well still generate more motivational force than the desire produced by her knowledge can generate.

For the individual who aims to pursue the right course of action, the only hope is that one's true knowledge can generate a sufficiently strong desire that will be forceful enough to counter the power of the passive emotions (IV P14). But even here, Spinoza cautions that "desire from the true knowledge of good and evil can be extinguished or checked by many other desires" if the former desire is not sufficiently proportionate in strength (IV P15). And it is this fundamentally disproportionate relationship between the

desire generated from adequate ideas of good or evil and the passive emotions arising from inadequate ideas that forms the core of Spinoza's theory of weakness of will. It is on the basis of this *core* relationship between knowledge and the emotions that Spinoza concludes, "I think I have thus demonstrated why men are motivated by uncritical belief (opinion) more than by true reasoning, and why the true knowledge of good and evil stirs up conflict in the mind and often yields to every kind of passion," adding, "hence the saying of the poet, [Ovid's description of Medea,] 'I see the better course and approve it, but I pursue the worse course.' Ecclesiastes seems to have had the same point in mind when he said: 'He who increaseth knowledge increaseth sorrow'" (IV P17; Ovid's *Metamorphoses*, VII, 20; *Ecclesiastes*, 1:18).

In this way, Spinoza thinks he has explained the relative power of inadequate knowledge over adequate knowledge, *and* the power of the emotions over true knowledge. Moreover, he has shown how weakness of will is not a 'problem' in human behavior or an exception in need of explanation. Rather, it is a product of our very natures as human beings. It is this explanation of weakness of will as the normal outcome of the power dynamics of our passive and active emotions being inadequate and adequate ideas, that is the backbone of Spinoza's approach to weakness of will.

Understanding Spinoza's theory in terms of its causal mechanisms has allowed me to argue that the first branch of Spinoza's theory can and does carry the weight of his basic theory of weakness of will. In addition, I have further suggested that Spinoza's theory of the emotions is 'successful,' insofar as it can provide an account of weak-willed behavior. Spinoza himself concludes his account by stating, "I have thus briefly explained the causes of human weakness and inconstancy, and why men do not abide by the precepts of reason" (IV P18, *Scholium*). At the same time, the account of weakness of will provides a compelling

illustration of his broader theory regarding the power of knowledge and the emotions in determining our everyday actions, emphasizing the power of knowledge as much as its limits compared with the external powers we are acted upon. Spinoza, having rejected the notion of an autonomous faculty of the will, sees human beings as parts of nature who can obtain some degree of freedom, but who will never be absolutely free.

5. R.M. Hare's Prescriptivist Account of Weakness of Will

Spinoza approaches the problem of weakness of will as part of his effort to deny the existence of an independent faculty of the will. In the 20th century, R.M. Hare turns to the issue because it is frequently presented as an objection to his prescriptivist account of moral judgment.

There is interpretive debate regarding how well Hare succeeds in reconciling prescriptivism and the possibility of weakness of will. Some have argued, as Stroud and Tappolet do, that "Hare denies the possibility of weakness of will because he defends prescriptivism, which maintains that moral judgments like 'I ought to do *x*' entail imperatives" (2003, 2). Others have suggested that Hare is only able to resolve the problem because he redefines the phenomenon of weakness of will altogether (Matthews 1966). In this section, I suggest that Hare is able to accommodate both prescriptivism and weakness of will by appending additional physical and psychological faculties to his conception of evaluative judgments.

Hare defends a conception of prescriptivism that is perhaps most easily defined in contrast to its counterpart, descriptivism. The latter position suggests that a statement such as, 'I ought to walk the dog,' does not carry any special motivational force, and simply corresponds to a statement or proposition. By contrast, prescriptivism proposes that moral

terms (such as 'ought') and judgments (such as 'I ought to walk the dog') are understood as guides to action, and share some of the structure of imperatives ('Walk the dog!'). If you believe or state that you ought to walk the dog, this statement is structurally related to carrying out the action expressed in the statement.

Even more strongly, Hare holds that generally ethical statements of this sort were similar to imperatives that are universal in scope. In *Language of Morals* (1952, 20), he writes, "it is a tautology to say that we cannot sincerely assent to a command addressed to ourselves, and *at the same time* not perform it, if now is the occasion for performing it, and it is in our (physical and psychological) power to do so." In the same text, Hare goes so far as to propose a test to see whether an agent is really making an evaluative judgment or not, echoing the same prescriptive principle: "… the test, whether someone is using the judgment 'I ought to do X' as a value-judgment or not is, 'Does he or does he not recognize that if he assents to the judgment, he must also assent to the command 'Let me do X'?" (168-169). In this way, Hare holds that evaluative judgments are not merely descriptive, but have a motivational (or 'activation orienting') feature that is intrinsically related to an appropriate, corresponding action.

Perhaps unsurprisingly, however, Hare's prescriptivist position is vulnerable to criticisms regarding the possibility of weakness of will. If an evaluative judgment results in an appropriate action, how is weakness of will possible? Hare acknowledges the problem and notes that he is by no means the first philosopher whose commitments had drawn him into this trap. If evaluative judgments are intrinsically related to action, he remarks, then the "familiar 'Socratic paradox arises, in that it becomes analytic to say that everyone always does what he thinks he ought to (in the evaluative sense)" (11.2, 169). In his subsequent essay "Backsliding" in *Freedom and Reasons* from 1963, Hare aims to resolve this problem by

making logical room for weakness of will. He does so by returning to his definition of prescriptivism above ("It is a tautology…"), now emphasizing the caveat of physical or psychological capacity.

Hare wants to deny that special pleading for oneself, or what he calls straightforward 'hypocrisy,' constitutes weakness of will. Rather, he leaves open the possibility that an agent knows she ought to do something, but simply is not able to. Hare calls this something like a 'ought but can't' position, where a physical or psychological inability prevents the agents from realizing his or her evaluative judgment in the form of an appropriate action.

Hare doesn't discuss the material differences between physical and psychological incapacity, but he does specify that they have differing consequences in the context of weakness of will. Physical impossibility, which can also include a lack of knowledge or a lack of skills, means that the relevant imperative simply gets downgraded. If I know I should walk the dog but my leg is broken, I may feel some remorse about not walking the dog, but I also recognize that my practical circumstances reasonably prevent me from doing so. By contrast, psychological impossibility equally prevents an agent from carrying what he or she knows he or she ought to do, but Hare does not think the obligation is downgraded in the same way. Rather, the prescription is preserved in this case, Hare argues, resulting in a psychological circumstance best expressed by "curious metaphor of divided personality which, ever since this subject is first discussed, has seemed so natural" (1963, 81). Hare's prescriptivism is thus able to stand, but is negatively acted upon by a physical or psychological influence. There are in turn two possible ways to understand weakness of will: either an agent accepts a moral judgment, but is unable to obey it or, alternately, some part of the agent subscribes to this judgment, but another part does not.

In this way, although his discussion is rooted in linguistic analysis, Hare still arrives at the similar conclusion that weighing or evaluating what is better and worse plays a central in certain kinds of decisions. That is, he manages to preserve the evaluative, motivational force of moral judgments, and yet is still able to account for weakness of will.

6. Assessing Valuation-Based Accounts as Mechanisms

In this section, I use Craver and Darden's criteria to argue that Plato's Socrates, Spinoza, and Hare's valuation-based models are to varying degrees incomplete. From within his system, each author identifies 'gray boxes' (i.e., projected functions) pertaining to valuation, but these descriptions seriously underspecify the relevant sub-mechanisms. At the same time, I suggest that these accounts are to varying degrees compatible with existing empirical evidence.

6.1. Incompleteness

Ranking the three preceding valuation-based positions in terms of incompleteness (from most incomplete to least), I discuss the accounts in the following order: Hare, Plato's Socrates, and Spinoza.

Hare's account is so incomplete as to border on superficiality. While he puts forward a valuation-based account, his analysis only loosely gestures toward 'some' psychological incapacitation that would interfere with an agent's actions. It is good to know that there is room for one or more auxiliary psychological forces in an otherwise analytically related interaction between judgment and action; but what does 'psychological impossibility' really mean in the context of decision-making? Although Hare cites Ovid's Medea and St. Paul's admission in his letter to the Romans– at length, he argues, to capture the prescriptivist tone of the two passages – he does not enter into a more detailed, mechanical account of what

might be underway. For someone looking to develop a heuristic model for sifting through the empirical evidence, this position does not provide much to go on. Hare's account is strongly incomplete (Figure 3.1.).

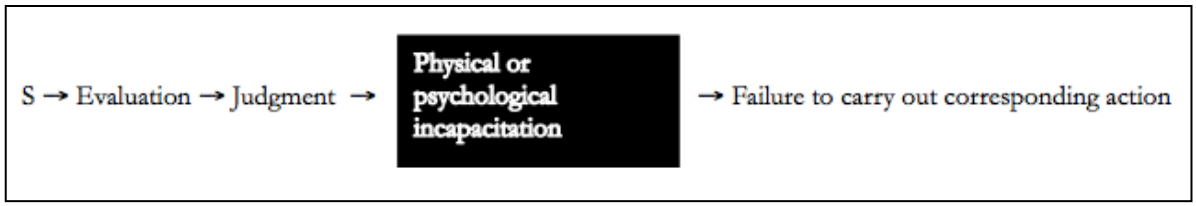

Figure 3.1. A representation of Hare's model of weakness of will. Using Craver and Darden's criteria, it is an incomplete mechanism schema (adapted from Craver and Darden, 2013, 87, Figure 6.1.).

In the *Protagoras*, Plato's Socrates uses the metaphor of weighing to describe the apparatus underlying weakness of will. At first glance, the very nature of a metaphor does not seem especially promising for characterizing mechanism schemas; after all, a metaphor is a symbol used to refer to an object or event, and it does not provide a literal description of it. In this sense, we have to characterize Socrates' account as painterly and mechanistically incomplete. But as we shall see in Section 5.2., Socrates' weighing metaphor captures a surprising number of the features of valuation that are consistent with recent empirical accounts.

Finally, Spinoza attempts to provide a detailed, causal account of weakness of will. The task of assessing its completeness depends on one's estimation of Spinoza's principle of parallelism.

In the *Ethics*, Spinoza explains that the idea of God is necessarily one and comprises both his essence and everything that necessarily follows from it. Since God's essence is expressed in an infinite number of attributes, each specific mode, or individual entity, is also expressed in infinite ways. However, Spinoza reasons that Extension and Thought are the only particular attributes of God which humans can know. This means that, for example, an

individual circle is expressed in infinite ways, but as human beings, we are only able to experience it in two ways: under the attribute of Extension, as an extended object existing in Nature, or under the attribute of Thought, as the idea of this existing object. Vitally, Spinoza's principle of parallelism emphasizes, "the order and connection of ideas is the same as the order and connection of things." This means that the circle existing in extension and the circle existing in thought is one and the same circle, but is simply expressed under two different aspects.

If – and for many, this is a big 'if' – one accepts the principle of parallelism, then Spinoza's discussion of the relative powers of knowledge may provide a developed systematization of weakness of will (Figure 3.2.). If one does not accept the metaphysical principle, however, Spinoza may provide a detailed description of the phenomenon, but one that may only reflect a strong internal coherence without explaining the mechanism 'as it really is.' I discuss the 'correctness' of Spinoza's account in Section 6.2. below.

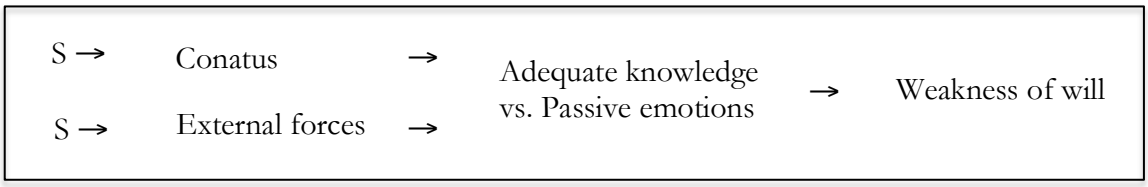

Figure 3.2. A representation of Spinoza's model of weakness of will. Using Craver and Darden's criteria, it is an incomplete mechanism schema (adapted from Craver and Darden, 2013, 87, Figure 6.1.).

6.2. Incorrectness

Hare's account of weakness of will is not sufficiently determined to test it against empirical evidence. The models of Plato's Socrates and Spinoza, however, can be discussed in light of contemporary empirical findings, and the consistencies can sometimes be surprising. In this section, I argue that Plato's Socrates and Spinoza's models correctly capture at least two out of these three aspects of valuation-based models:

1. Conceptions of value,
2. Measurement of value, and
3. Corresponding approach and withdraw behaviors.

At the same time, the main source of incorrectness in both of these accounts lies in their belief that valuation and decision-making rely on a single specialized mechanism. Contemporary findings strongly indicate that human beings rely on multiple interacting valuation systems. In Part II of the dissertation, I argue that these interactions can in turn cause suboptimal behaviors including weakness of will.

6.2.1. Value, Pleasure, and Pain

Depending on how one interprets the issue of hedonism in the *Protagoras*, Socrates may or may not conceive of pleasure and pain as indicators of value, or whether they are simply pursued in and for themselves. By contrast, Spinoza explicitly defines pleasure and pain in terms of indicators of value and wellbeing. In doing so, he is 'on the right track' in identifying value signals; nevertheless, contemporary neuroscience takes the further step of distinguishing between value signals and pleasure and pain signals.

On Spinoza's view, every individual's power of activity fluctuates over the course of both her everyday experiences and her life, according to her interactions with the surrounding environment. These fluctuations, in turn, constitute the root cause of emotional experience. For Spinoza, the emotions consist in "the affections of the body by which the body's power of activity is increased or diminished, assisted or checked, together with the ideas of these affections" (*Ethics* III, D3). Spinoza argues that fluctuations in the body's power of activity are paralleled in the mind's capacity to think. As a result, the constitutions of individual emotions are determined by increases and decreases in an individual's power of activity, and his or her associated power of thought. Spinoza explains, "we see then that the

mind can undergo considerable changes, and can pass now to a state of greater perfection, now to one of less perfection, and it is these passive transitions (passiones) that explicate for us the emotions of Pleasure (leatitia) and Pain (tristitia)" (*Ethics* III, P11, Sch.). 'Pleasure' consists in the "passive transition of the mind to a state of greater perfection," while 'pain' corresponds to the "passive transition of the mind to a state of less perfection" (*Ethics* III, P11, Sch.). When the mind is conscious of its conatus, it experiences what Spinoza calls 'will,' and he identifies these three emotions – pleasure, pain and desire – as the three primary emotions of human experience.

Contemporary neuroscience similarly identifies sub-mechanisms for signaling value, but distinguishes between value, pleasure, and pain. Notably, Redish argues that while people often suggest that 'people seek pleasure and avoid pain,' it would be more accurate to write, "people seek things that they recognize will have high value" (2013, 23). It is now known that value, pleasure and pain are dissociable and regulated by different neurochemical systems in the brain. There are at least three separate, interacting mechanisms (Figure 3.3.).

	System		
Stimulus	Evaluation	Consequence	Anticipation/response upon lack of delivery
Positively valenced	Pleasure/euphoria	Reinforcement	Disappointment
Negatively valenced	Pain/dysphoria	Aversion	Relief

Figure 3.3. The object provided and the three corresponding systems relating pleasure, value, and expectation. (Adopted, with modifications, from Redish 2013, 32, Figure 4.1.).

The mechanism responsible for experiencing pleasure and displeasure, or 'euphoria' and 'dysphoria,' corresponds to the opioid receptor system in the brain. The system responsible for experiencing value is known as the 'reinforcement' and 'aversion' system and enables an agent to learn which entities and experiences are worth pursuing, and which should be avoided in the future. This system is governed by the dopamine system in the

brain, and will be the main focus of my mechanistic explanation of weakness of will in Chapters 4 and 5. Finally, Redish argues for the necessity of a distinct mechanism that recognizes change, which he calls the "disappointment" system, which allows an agent to anticipate certain rewards and thus either be unsurprised, disappointed, or relieved when a certain expected outcome does not occur; the neural and chemical system underlying this last system is somewhat less clear. As noted in Figure 16, there are also parallel systems for pain, aversion, and relief, but these systems are currently even less well understood.

In this way, Spinoza's mechanism may be said to correctly identify the need for valuation signaling, but does not quite finesse the distinction between value and pleasure and pain to the same degree that contemporary neuroscience is able to.

6.2.2. Measuring Value

Despite attributing potentially diverging functions to the experiences of pleasure and pain, both Plato's Socrates and Spinoza underscore an agent's need to measure value, i.e., to be able to quantify and compare what benefits and harms us as living organisms. In the *Protagoras,* this principle of measurement lies at the heart of Socrates' weighing metaphor as well as his two fundamental principles of practical reasoning and choice:

> (L1) "If you weigh pleasant things against pleasant, the greater and the more must always be taken; if painful things against painful, the fewer and the smaller," and
>
> (L2) "If you weigh pleasant things against painful, and the painful is exceeded by the pleasant – whether the near by the remote or the remote by the near – you have to perform that action in which the pleasant prevails; on the other hand, if the pleasant is exceeded by the painful, you have to refrain from doing that" (356b-c).

Similarly, Spinoza returns to the concept of conatus to explain the quantification of value, discussing "the nature and strength of the emotions" (*Ethics* III, Preface). On his view, the

mind is necessarily conscious of its conatus, he explains, and "whatsoever increases or diminishes, assists or checks, the power of activity of our body, the idea of said thing increases or diminishes, assists or check the power of thought of our mind" (*Ethics* III, P9, P11.) Correspondingly, when this affection of the body increases or decreases his or her *vis existendi*, it is mirrored by an idea of this increase or decrease, which, descriptively, he or she experiences as a certain intensity of emotion, namely, of the emotions of pleasure or pain.

This notion of valuation is now well supported in the literature. For example, Damasio argues that the basic goal of all biological organisms is "healthy survival to an age compatible with reproductive success" (2010, 48). This basic goal, in turn, informs how organisms ascribe values to the entities and events they interact with, in two distinct ways. First, organisms ascribe values in relation to their 'general maintenance' routine, which they perform in order to stay within their healthy homeostatic range. Second, they also ascribe values in relation to what Damasio calls 'particular regulation,' that is, in relation to those processes which organisms undertake in response to their changing environments. Damasio explains that the "continuous representation of chemical parameters within the brain allows nonconscious brain devices to *detect and measure* departures from the homeostatic range and thus act as sensors for the degree of internal need. In turn, the measured departure from homeostatic range allows yet other brain devices to command corrective actions and even to promote *incentive* or *disincentive* for corrections, depending on the urgency of response" (2010, 49). In this way, organisms systematically register and 'evaluate' their interactions with their environments, and ascribe variously positive and negative values to the many entities they interact with.

In brains that are capable of representing internal states in the form of brain 'maps,' changes in these ongoing assessments of the homeostatic range are also experienced as

conscious representations. Significant changes in the organism's body states trigger evolved emotional programs, causing it to execute a certain sequence of appropriate, responsive actions, and equally leading it to experience a set of certain cognitive states. These cognitive states correspond to what are then consciously experienced as feelings, or the "composite perceptions of (1) a particular state of the body, during actual or simulated emotion, and (2) a state of altered cognitive resources and a deployment of certain mental scripts" (Damasio 2010, 116).

The specific computational and dopaminergic systems for measuring value will be discussed at length in Chapters 4 and 5.

6.2.3. Approach and Withdraw Behaviors

Finally, both Plato's Socrates and Spinoza describe approach and withdraw behaviors that are consistent with contemporary scientific models. Socrates specifies that human beings naturally choose what is more pleasant and less painful, and seek out what is pleasant and avoid what is painful (356b-c). Similarly, Spinoza examines how feelings of pleasure, desire, and pain, as well as of love and hatred, motivate our basic behaviors. At the most essential level, he maintains that we endeavor to affirm whatever causes us pleasure or which we imagine causes us pleasure. By contrast, we endeavor to avoid whatever causes us pain, or what we imagine causes us pain (*Ethics, III* P25, 26, 28, 36).

Although these characterizations are relatively intuitive, approach and withdrawal behaviors are fundamental to all living organisms. As psychologist Andrew Elliot notes, "both approach and avoidance motivation are integral to successful adaption; avoidance motivation facilitates surviving, while approach motivation facilitates thriving" (2006). Even the most basic organisms exhibit approach and withdraw behaviors, with more complex

organisms generally possessing higher numbers of approach-withdraw mechanisms (Tooby and Cosmides 1990). Without them, any processes of perception, valuation or even deliberation would be ineffectual. Because of its central role, the approach – withdrawal distinction is one of the important concepts in behavioral psychology and, in turn, in computational neuroscience (see also Huys *et al.* 2011, Cools *et al.* 2010).

6.2.4. Reliance on a Single System

The main source of incorrectness in both of the accounts of Plato's Socrates and of Spinoza lies in their belief that valuation and decision-making rely on a single specialized mechanism. Contemporary findings strongly indicate that human beings rely on multiple interacting valuation systems. In Part II of the dissertation, I will argue that these interactions can in turn cause suboptimal behaviors such as weakness of will.

PART II

LEARNING FROM MACHINES, ANIMALS, AND HUMANS BEINGS: A NEW VALUATION-BASED THEORY OF DECISION-MAKING AND WEAKNESS OF WILL

CHAPTER 4

ADOPTING A DIFFERENT STARTING POINT: COMPUTATION-BASED APPROACHES TO DECISION-MAKING AND WEAKNESS OF WILL

1. Introduction

In Part I of the dissertation, I argued that philosophers have presented two main mechanism schemas for understanding weakness of will, offering either 'syllogism-based' or 'valuation-based' models of the phenomenon. The former correspond to models based on the structure of the logical syllogism. The latter emphasize the processes of valuation, that is, the processes whereby we come to value and seek out what benefits us as living organisms.

Although syllogism-based models have historically dominated philosophical treatments of the issue and remain influential in contemporary philosophy, they contain significant explanatory gaps and are thoroughly inconsistent with contemporary empirical findings in psychology and neuroscience. Valuation-based models of practical reasoning and weakness of will have historically proven to be less prevalent in the philosophical literature, and also suffer from key explanatory gaps; however, they have a preliminary grasp of several important aspects of decision-making as understood today.

Part II of the dissertation renews the philosophical tradition of searching for the mechanism underlying weakness of will. In particular, it aims to discover a mechanism schema that is consistent with contemporary behavioral psychology and computational neuroscience. Based on converging evidence from computer science, psychology, and neuroscience, I argue that the human brain employs three dissociable mechanisms to make choices. The 'Pavlovian' mechanism corresponds to hard-wired approach and withdrawal responses. 'Goal-directed' behaviors map out different options and assess them in light of

specific goals. 'Habit-based' behaviors learn the value of actions over time and choose the most consistently valuable option. Suboptimal interactions between these three decision-making mechanisms in turn generate two different categories of weakness of will, which are etiologically but not psychologically distinguishable.

In Section 2, I use the example of a robot named 'AL1C3' (pronounced 'Alice') to outline four basic features that are needed to generate 'value,' and thereby enable an agent to make and act on complex decisions. These four features are as follows: an agent must 1) be capable of possessing a goal or set of goals 2) have the basic capacity to monitor and regulate itself 3) possess a mechanism or set of mechanisms that enable(s) her to assign values to experienced or imagined objects, actions, and events, and, finally, 4) possess a mechanism or set of mechanisms that enable(s) her to predict the value of several competing alternatives as reliably as possible, thus allowing her to select and execute a coherent course of action. Each of these umbrella features contains multiple sub-mechanisms. If successfully integrated, they should enable an agent to navigate between competing courses of action. They should also help explain why it is possible to behave in a weak-willed manner.

Next, I turn to principles from reinforcement learning to provide a normative account of how these four capacities can be worked out in detail, particularly focusing on capacities 3) and 4), namely, attributing values and predicting values. In Section 3, I provide a brief historical and theoretical sketch of the field of reinforcement learning. In Section 4, I discuss the specific theoretical and computational underpinnings of the Pavlovian, goal-directed, and habit-based decision-making mechanisms.

In Section 5, I analyze how these mechanisms interact in both complementary and competitive ways (Daw *et al.* 2005, Daw *et al.* 2006, Dayan 2008, Huys *et al.* 2008, Redish 2013). Based on my analysis, I argue that suboptimal interactions between these three

decision-making mechanisms generate two general categories of weakness of will. I outline two detailed hypotheses regarding how weakness of will is generated.[32]

The first hypothesis states that weakness of will is sometimes the product of competition between the goal-directed and habit-based controllers, which generates a weak and remediable suboptimal action response.

The second hypothesis states that, under different circumstances, weakness of will is the product of competition between the Pavlovian and goal-directed controllers, which generates a more robust, non-remediable suboptimal action response. As part of the second hypothesis, I further propose that there are two Pavlovian-based types of weakness of will, a 'Pavlovian-Cognitive' weakness of will and a 'Pavlovian Behavioral' weakness of will. Pavlovian Cognitive weakness of will is characterized by suboptimal cognitive pruning of a decision tree. Pavlovian Behavioral weakness of will refers to a Pavlovian approach or withdrawal behavior that hinders optimal action-selection.

Finally, in Section 6, I propose that although my mechanism schema shares some of the features of a valuation-based model, three central differences distinguish it from both the historical syllogism- and valuation-based accounts presented in Part I, Chapters 2 and 3. First, my account moves beyond the longstanding philosophical assumption that there is a single mechanism underlying practical reasoning, and adopts a view that is compatible with current scientific opinion, namely, that there are three mechanisms underlying human decision-making. Second, it breaks with the traditional conception that weakness of will itself is a single phenomenon and proposes that it is instead better understood as a constellation of behaviors. This pluralistic position helps explain some of the long-standing disagreements surrounding the issue in the history of philosophy.

[32] It is likely that there are additional kinds of weakness of will. See Chapter 5, Section 2.2. on Other Vulnerabilities.

Third, I suggest that a key difference and advantage of my account lies in its ability to use Bayesian model-fitting techniques to verify the degree to which an agent truly 'knows' the detrimental consequences of her actions in different types and instances of weakness of will.

In Chapter 5, I return to Craver and Darden's criteria for assessing mechanism schemas and aim to evaluate my own proposal. I particularly focus on the 'vices' of incompleteness and incorrectness. Regarding incompleteness, I discuss two areas of my account that I believe correspond to explanatory 'gray boxes,' namely, the as-yet underdetermined mechanism of 'arbitration' (Daw *et al.* 2005), and the unaddressed, additional sources of 'vulnerability' described in Redish's (2013) account of the three decision-making mechanisms. To evaluate the relative correctness or incorrectness of my account, I turn to behavioral and neuroscientific evidence suggesting that multiple decision making systems are indeed operational in animals and, especially, in human beings.

But first, let's take a step back and ask a few basic questions about what it means to 'make a decision.'

2. Some Basic Requirements for Decision-Making and Weakness of Will

In Chapter 2, I argued that syllogism-based accounts of weakness of will are a) theoretically incomplete and b) inconsistent with recent empirical findings regarding how human beings evaluate, judge, act, and fail to act in real-life circumstances. What, then, must be true for an agent to make decisions and, occasionally, to fail to act on what she has decided to do? What if I wanted to design a robot named AL1C3 (pronounced 'Alice') such that it would be capable of making semi-independent decisions in the field?

Neuroscientist David Redish (2013, 9-10) uses the example of a thermostat to outline

three central components of decision-making. A thermostat uses negative feedback to regulate the temperature in a house, depending on whether it is too hot or too cold inside. To do this, Redish argues, the thermostat needs to a) perceive the world, i.e. detect the temperature b) determine what needs to be done, i.e., compare the temperature to the set point and determine whether it needs to increase or decrease the heat, and finally, c) take action, i.e., either increase or decrease the heat.

Using the example of AL1C3, I suggest that an additional component is required for decision-making, namely, a decision-maker must possess a goal or set of goals. In Redish's example, for instance, the thermostat is first and foremost defined by its purpose of regulating the temperature; the three additional components of decision-making follow from this defining function. With this in mind, I argue that four basic features are needed for an agent to make and act on complex decisions. An agent must a) be capable of possessing a goal or set of goals, b) have the basic capacity to monitor and regulate itself, c) possess a mechanism or set of mechanisms that enable(s) her to assign values to experienced (or imagined?) objects, actions, and events, and, finally, d) possess a mechanism or set of mechanisms that enable(s) her to predict the value of several competing alternatives as reliably as possible, thus allowing it her to select and execute a coherent course of action. AL1C3 must thus possess at least these four basic abilities in order to realize its objective.

2.1. Goals

The first capacity AL1C3 will need is a general *goal* or set of goals. It should have objectives that it achieves in a variety of ways, thereby making possible specific, intermediate options and decisions. If it has a goal, AL1C3 can set about its business and evaluate how well it is

doing relative to its task at each step.

Even in an entity that is indifferent to the concept of survival, goals are essential to an efficient mechanism because they provide a baseline for assessment or 'feedback.' In its simplest terms, a goal can be characterized as a desired state (Montague 2006, 53). This can be as simple as property of an electrical circuit or a physical brain state resulting from eating food or having sex, or as complex as paying taxes or planning a career.

Although we could ask what it means for a machine to possess a goal, computational neuroscientist Read Montague argues that what is actually the most interesting about goals in either machines or brains is how they can be translated from relatively abstract and complex goals into simple, realizable commands. From the perspective of efficiency, it will be important not to frame goals in terms of an endless series of specific commands, but rather by providing indirect feedback. Montague remarks, "indirect is the key word here. Specifying complex goals directly is not an efficient way to build goals into a system. [For example,] our nervous system sidesteps the difficulty of directly defining complex goals. Instead, it comes equipped with a series of guidance systems" (2006, 49). Imagining a game of 'hot or cold' can be useful for understanding how these guidance signals work.

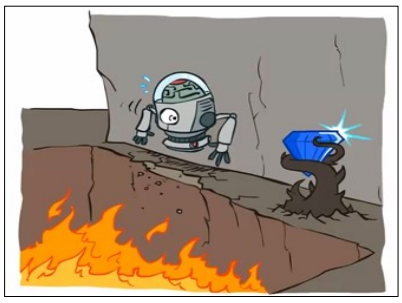

Figure 4.1. A self-guiding, problem-solving robot such as AL1C3 (Image from Abbeel & Klein, 2013).

Pretend you are performing a training exercise with a machine. You have 'hidden' a reward at the far ends of its operating state. You start out by placing it randomly in the room, and you provide a series of cues, depending on how well it is doing relative to the reward (e.g., 'better,' 'worse,' 'absolutely on the wrong track,' 'you are getting close!', etc.).

What you *don't* do is give the robot a list of a hundred explicit 'don't' commands (e.g., 'Don't look under the table,' 'don't look under the lamp,' 'don't look on the piano,' 'don't look under the cushion,'etc.) or a set of detailed, step-by-step commands for how to get to the candy bonanza (e.g., 'Put left wheel forward,' 'put right wheel forward,' 'raise arm,' etc.).[33] This would be a less efficient or useful way to seek out the reward – it would take forever, and a whole new set of rules would need to be drawn up each time. Along very similar lines, Montague argues, for either machines or living organisms, indirect guidance signals are far a more effective means for realizing complicated tasks in new and ever-changing circumstances.

In this way, if it is to be more than a remote-controlled vehicle, AL1C3 must not simply be commanded 'left-right, left-right' at each step (Figure 4.1.).[34] *Goals* are key features of decision-making by providing an opportunity and an ultimate baseline for assessment or 'feedback' (see also Sutton and Barto on goals, 1998, 5)

2.2. Basic Regulation

Next, AL1C3 must be able track how well it is doing along two separate axes. First, it must be able to execute a minimum maintenance routine to ensure that it can go on carrying out its tasks. This can be called 'basic regulation,' and loosely corresponds to a set of physical considerations that AL1C3 must take into account if it is to survive through its projected lifespan. For example, AL1C3 must monitor its battery life and its CPU temperature. If it fails to regulate one of these, it may very well end its mission.

---

[33] Montague uses the example of playing 'warmer, colder' with young children, but this occasionally causes confusion as to whether he is talking about machine learning or animal learning. My example is perhaps less illustrative, but should help the reader stick to machine learning for now.

[34] Images by Dan Klein, from Peter Abeel's CS188 Artificial Intelligence (Spring 2013) at the University of California, Berkeley.

2.3. Specific Regulation

Second, AL1C3 needs to be able to track how well it is doing in relation to realizing its more specific goal or set of goals. Using the example of searching for a rock sample, AL1C3 needs to know if it is doing 'better' or 'worse' in relation to this goal. For this reason, it needs to be able to mark the entities and events it experiences, depending on whether they aided or hindered the realization of its goal, e.g., 'This crater was detrimental,' or 'This pile of rocks brought me closer.' This capacity can be called 'value attribution,' and we will see how this mechanism might work in Section 3.[35]For now, it is enough to recognize that an entity must have some way of assessing or evaluating its environment if it is to make any decisions about its actions.

2.4.Value-Based Decision-Making: Choice

Finally, AL1C3 needs a means to weigh and 'decide' between multiple different courses of action. This would likely mean having one or both of the following two options: 1) having multiple task-specific systems, which intermittently interact, and /or 2) possessing a 'common denominator unit' capable of i) receiving and comparing competing inputs and ii) generating relatively coherent action outputs for a range of different tasks.
Determining how this capacity might work will hold the key to understanding weakness of the will.

Of course, AL1C3's model needs to be fleshed out in much more detail. This can be done in two ways: first, by turning to computational modeling, which can provide a

[35] Although robots, like evolved organisms, can come preprogrammed with certain 'hardwired' behavioral responses (e.g. approach/withdrawal behaviors) and somewhat more flexible predispositions, active value attribution is based on learning retrospectively from experience.

normative account of optimized decision-making[36], and second, by looking to descriptive behavioral and neuroscientific research, which can suggest how these mechanisms are instantiated in living organisms.

Before turning to findings from the computational theory of reinforcement learning, it will be useful to say a few introductory words about the field of reinforcement learning itself.

3. The Historical and Theoretical Background of Reinforcement Learning

In the first half of the twentieth century, behaviorism introduced the first systematization of psychology and set rigorous standards for what was acceptable in psychological experiments. Although subsequently eclipsed by the so-called 'Cognitive Revolution,' behaviorism had at least one major consequence in that it gave rise to the field of 'mathematical psychology,' which sought to use statistics to understand learning processes. In particular, based on early research by Bush and Mosteller (1951) and Kamin (1969), the Rescorla-Wagner (1979) model emerged as a highly influential account of animal learning. The model advanced the idea that "learning occurs only when events *violate expectations*," and

---

36 Importantly, in Reinforcement Learning, Sutton and Barto (1998) emphasize that reinforcement learning examines optimized or idealized models of learning. Similarly, Niv and Montague (2009) describe reinforcement learning as a normative framework for analyzing a specific kind of learning. As they put it, reinforcement -learning models calculate "a means by which optimal prediction and action selection can be achieved, and exposes explicitly the computations that must be realized in the service of these. In contrast to descriptive models that describe behavior as it is, normative models study behavior from the point of view of its hypothesized function – that is, they study behavior as it should be if it were to accomplish specific goals in an optimal way" (2009, 332).

On Niv and Montague's view, adopting this normative or idealized approach has two advantages. First, if one views evolved behavior as a nearly optimal adaptation, then reinforcement learning can generate testable hypotheses regarding animal behavior. Second, discrepancies between predicted optimal behavior and actual, sub-optimal behavior can help scientists understand the constraints defining behavior, or perhaps suggest that the agents are optimizing behavior in a previously unrecognized way, as has been successfully shown in behavioral economics (Kahneman and Tversky, 1982; Gigerenzer *et al.* 1992). In either case, it is useful to know the optimal solution to specific decision-making scenarios, even if real-life models turn out to operate sub-optimally.

this key principle was quickly able to model several previously perplexing features of animal behavior (Niv and Montague, 2008, 333). In particular, the Rescorla-Wagner rule provides a consistent account of classical conditioning behaviors. In the mid-1990s, Sutton and Barto then adapted the Rescorla-Wagner model to examine how the expectation-violating experiences of trial and error allow an agent to learn to make efficient decisions over time (discussed in Section 4.1. below on Instrumental Decision-Making).[37]

In their research, Sutton and Barto examined a set of models known as Markov Decision Processes (MDPs), named after Russian mathematician Andrei Markov (1856-1922). MDPs provide a mathematical framework for modeling decision making in situations where an agent can control its actions, but cannot control the outcomes of those actions. To illustrate the process, we can return to the example of AL1C3, set down in an environment with a diamond (positive reward), a fire pit (negative reward), and a wall blocking its path (obstacle) (Figure 4.2.).

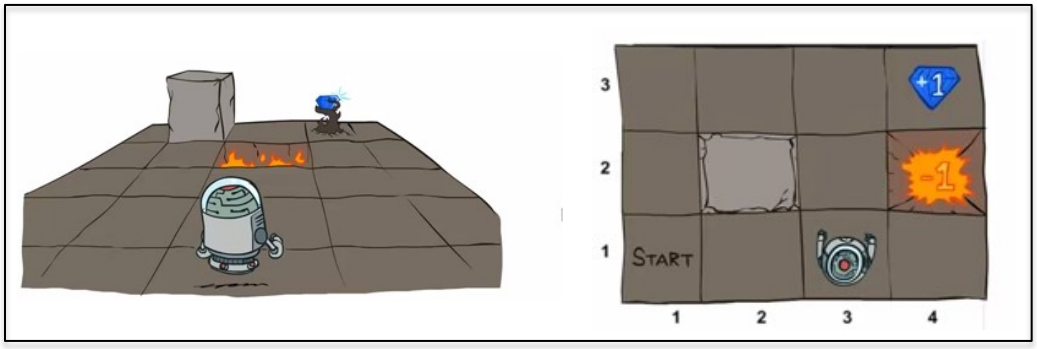

Figure 4.2. Left: AL1C3 in a new environment. Right: AL1C3's new environment formalized into a grid space (Image from Abbeel & Klein, 2013).

AL1C3's goal is to maximize its rewards, but it cannot directly determine where its actions will take it (Figure 4.3.). This is because although AL1C3 may choose to head north, it may have only an 80% success rate, with its actions resulting in an eastward or westward move

[37] For a more detailed account, see also Barto (1995).

the remaining 20% of the time. Or in other words, AL1C3 follows an intermediate, internal directive to move north, but only realizes this intermediary step 80% of the time (in this grid space). In turn, these intermediate successes and failures have an impact on the AL1C3's overarching goal of reaching the diamond and maximizing its rewards.

To make matters worse, AL1C3 is faced with what is known as a 'living reward,' meaning that it costs it, e.g., -0.2, for every transition it makes. This means that AL1C3 must try to reach the diamond as quickly as possible without landing in the fire pit. As such, the defining problem of MDPs is, 'what is the most efficient path or *policy* for an agent to adopt in order to achieve its goal?'

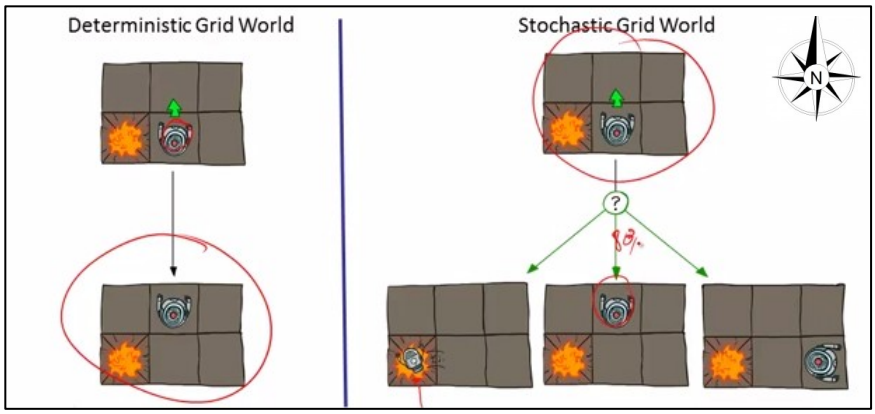

Figure 4.3. Left: In some types of reinforcement learning, AL1C3 can simulate movement and learning in a deterministic environment. In a determined grid-space, AL1C3 could 'choose' to move north (up), as illustrated by the green arrow, (top left), and would successfully move north, circled in red, 100% of the time (bottom left).
Right: By contrast, in an MDP, AL1C3 can 'choose' her actions, e.g. to move north (up) (top right), circled in red, but the outcome of her actions is stochastic (bottom right). There is an 80% chance AL1C3 will move according to plan, but there is also a 10% it will move east (to the right) and a 10% chance it will move west (to the left), and in this case, right into the fire (Image from Abbeel & Klein, 2013).

An MDP is formally defined by: a set of states ($s \in S$); a set of actions ($a \in A$), a transition function reflecting the probability that taking action *a* in state s leads to s^1, i.e., ($T(s, a, s^1) = G$, where the transition function takes ordered triples of state, action, state, and gives a probability 'G' that taking action 'a' when in state 's' leads to being in state s^1; an

immediate reward function ($R(s, a, s^1)$) = N, where the reward is received after each transition to state s^1 from state s, where 'N' stands for a number, and higher numbers correspond to greater rewards; a starting state; and sometimes a terminal state (Russell and Norvig, 2009). The agent's goal is to find an optimal policy, P*, that allows it to maximize its cumulative rewards, i.e., the sum of the individual rewards it encounters in the different states. In our example, maximizing the cumulative reward involves a policy that avoids the fire pit, reaches the diamond, and keeps the cost of the 'living reward' low by keeping the number of transitions as low as possible.

In their work, Sutton and Barto focus on a special set of MDPs, where both the transition probabilities and rewards in each state are initially unknown to the agent. The members of this special subset of MDPs are known as reinforcement learning. As a special version of an MDP, basic reinforcement learning models consist of four key components: a set of states, a set of actions, rules governing the transitioning between states, and rules that determine the immediate reward of a transition. However, in these setups, the machine does not *know* how the transitions work, or what the rewards will be in each state. Rather, it must learn what it learns from experience and try to predict the values of transitions and rewards that it has not yet made (Figure 4.4.). To do so, the machine can be designed to deploy one or more computational strategies to try and best estimate the value of different courses of action,

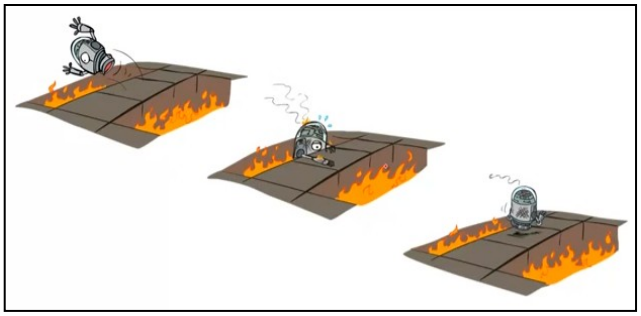

Figure 4.4. In reinforcement learning, AL1C3 does not know the transition rules or rewards in advance. It must learn using trial and error (Image from Abbeel & Klein,

and ultimately to arrive at a policy that will allow it to collect as much reward as possible.[38] The task of elucidating the most effective computational strategies has been the major task and achievement of reinforcement learning. Referring to the complexities involved in analyzing choice, computational neuroscientist Peter Dayan remarks,

> a seemingly obvious way to formalize choice is to evaluate the predicted costs and benefits of each option and pick the best. However, seething beneath the surface of this bland dictate lies a host of questions about such things as a common currency by which to capture the costs and benefits, the different mechanisms by which these predictions may be made, the different information the predictors might use to assess the costs and benefits, the possibility of choosing when or how quickly to act as well as what to do, and different prior expectations that may be brought to bear in that vast majority of cases when aspects of the problem remain uncertain (in Engel and Singer, 2008, 51).

In the face of these many challenges, however, researchers have succeeded in elucidating at least three computational mechanisms, known as the (conditioning) Pavlovian and (instrumental) goal-directed and habit-based systems. The Pavlovian system corresponds to automatic response behaviors. The goal-directed system comprehensively maps out different options and assesses them in light of specific goals. The habit-based system learns the value of actions in specific states over time and chooses the most consistently valuable option. I discuss each of these systems in detail in Section 4. In Section 5, I go on to hypothesize how interactions between these systems may elicit weakness of will.

4. The Theoretical and Computational Underpinnings of the Three Controllers

Evidence from computational modeling suggests that decision-making behaviors may broadly be divided into two categories: classical and instrumental conditioning. The two

[38] In 'supervised learning,' a machine learns by having an external supervisor present it with examples and by then using these examples to infer appropriate functions or actions (Mohri et al. 2012). Examples consist of correct input-output pairs that allow the machine to learn the correct action policy, and if the machine calculates a sub-optimal action policy, the external supervisor can provide it with corrective signals. By contrast, reinforcement learning takes place without pre-existing examples or a supervisor's corrections. In this kind of learning, the machine proceeds alone by trial and error and gradually generates its action policies by interacting with the environment over time.

categories of learning behaviors are distinguished by the relative importance or non-importance of action on the part of the agent. Classical conditioning will correspond to the Pavlovian decision-making mechanism. Instrumental conditioning will correspond to the goal-directed and habit-based decision-making mechanisms.

On the one hand, classical conditioning refers to cases where an agent learns that a certain stimulus predicts a positively or negatively significant outcome, but where this relationship is independent of any action or behavior on the part of this agent. For example, a robot may learn, "when song, then reward," or, "when song, then loss." But this relationship should hold independently of whether the robot acts or doesn't act. By contrast, operant or instrumental conditioning involves an action on the part of the agent. For example, a robot may learn that if it presses a lever, it gets a reward or a loss. Here, the outcome is made contingent on the robot's completion of the relevant action (Dayan *et al.* 2006). The difference between classical and instrumental conditioning thus depends on the relative role of action.

For the purposes of this chapter, it will be useful to remember that the Pavlovian decision-making mechanism is an instance of classical conditioning, i.e., as an action-independent behavior, whereas 'goal-directed' and 'habit-based' behaviors are two types of action-contingent or instrumental behaviors.

In each of the sub-sections below, I turn to the reinforcement learning literature to describe the computational mechanisms characterizing the Pavlovian, goal-directed and habit-based decision-making systems.

4.1. Instrumental Decision-Making

In natural settings, rewards and punishments typically come as the result of an action or

set of actions undertaken by an agent. To this end, it may be said that instrumental behaviors are fundamental to the processes of learning, decision-making, and action selection. In particular, agents develop courses of action, or *policies*, to increase their intake of rewards. There are two main ways agents learn to develop appropriate policies, namely, goal-directed and habit-based learning.

4.1.1. The Goal-Directed (Model-Based) Decision-Making System

Goal-directed learning aims to maximize an agent's rewards while minimizing its exposure to unknown experiences in the real world. In order to do so, it learns about some aspects of its environment and then simulates what its replica, AL1C3*, would do in a model version of its circumstances (Figure 5). For this reason, goal-directed learning is also frequently called 'model-based' learning.

More formally, in goal-directed learning, AL1C3 proceeds in three stages. First, it undergoes a 'training' period, where it collects experiences from its real-life environment. Second, it uses the information it has gathered to build a model of itself in its environment. In particular, since it is an incomplete MDP (i.e., a standard reinforcement learning problem), it tries to solve for the values it does not know, namely, the transition function ($T(s, a, s^{1})$) = G and the reward function ($R(s, a, s^{1})$) = N. By solving for these values, AL1C3 can arrive at an estimate-based model of its situation. Thirdly, it can use this model to try and calculate the best policy for negotiating its model environment.

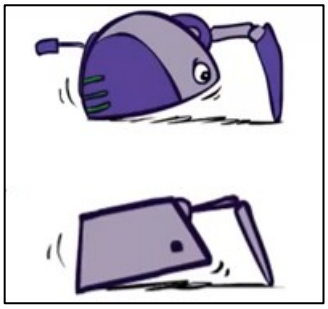

Figure 4.5. AL1C3 floats a model of itself, AL1C3*, to try and plan an optimal action policy (Image from Abbeel & Klein, 2013).

In the literature, this three-step process is sometimes represented as a decision-making tree, which lends this strategy its alternate title of "tree search." In order to picture AL1C3 as completing a tree search, we can imagine that we have set it up before a 'choose your own adventure' game.[39] In the game, AL1C3 takes on the role of the protagonist, e.g. a knight in shining armor, and makes choices regarding the knight's actions that ultimately determine the outcome of the adventure. In 'choose your own adventure' games, it is often up the player – in this case AL1C3 – to decide whether to be the 'good guy' and choose helpful actions at each successive junction, or to be the 'bad guy' and choose selfish actions instead. By analogy, model-based decision-making involves making choices that will arrive at the most rewards at each individual state, and hence secure as much value overall as is possible. What makes model-based learning distinctive is that, faced with this task, AL1C3 would try and model all of the possible branches of the adventure in advance, and then only pursue the one it has projected as being the most rewarding. In particular, ALIC3 would use previous experience, either on the same problem or similar problems, to try and figure out the best course of action.[40]

Goal-directed learning is also called 'model-based' learning and 'tree search.'

Box 1. What's in a name?

[39] As another way to illustrated model-based learning, Peter Dayan uses the example of Gary Kasparov playing chess with Deep Blue. At a particular position in a game of chess, Kasparov may be presented with three or four moves that are immediately possible, which subsequently branch into dozens of possible further courses of action. The reason this case is so illuminating is because, as Dayan puts it, "the trouble in chess is that the branching factor, that is to say, the number of moves at any one time, is of the order of 30… the size of this tree is absolutely, monumentally vast" (2011). In this regard, planning a model for all the possible moves in a game of chess is a fruitful metaphor for planning all the possible moves in other complex situations, insofar as it demonstrates just how challenging developing and using such a model would be.

[40] Model-based learning is based on the following recursive definition ('recursive' because it takes each subsequent state into account): $Q(s,a) = R(s) + \Sigma_{s'} T(s, a, s') \max_{a'}[Q(s', a')]$, which can roughly be translated as: overall value attained = (reward at state s) + sum of all successor states ((transition function)(best action in successor states)). In model-based learning, an agent simply estimates *R* and *T*, and then adde up the anticipated individual reward associated with each successive state. This process is called value iteration (Balleine*et al.* 2009). To stay with the chess metaphor, a novice player may explicitly try to work out the costs

Given the complexity of what AL1C3 faces, the strengths and weaknesses of goal-directed learning quickly become apparent. Its advantage is that it is situationally flexible. AL1C3 can build a new model or tree at any stage of its decision-making process, such that, if and when its circumstances change, the model can change as well. This means it is worth performing a tree search in novel, high-stakes situations. In addition, AL1C3 does not need very many samples to develop a model, because it makes short-term predictions about the immediate consequences of an action or set of actions. This renders it statistically efficient and reduces the probability of making an unexpected, potentially fatal error.[41] For example, if AL1C3 is navigating the grid space depicted in Figure 3, the goal-directed mechanism ensures that AL1C3 computes something like, 'If I fall into the fire pit, then the consequences will be very negative indeed.' This is in direct contrast to what AL1C3's 'thought process' would be using the habit-based controller. This would have AL1C3 simply wander around the grid space and gradually learn by trial and error (see Section 4.1.2. on Habit-Based Decision Making below). If we imagine interactions with the fire pit on a trial-and-error basis, we can see how goal-directed learning can help minimize high-cost mistakes.

Unfortunately, goal-directed learning is potentially very computationally challenging, because it requires putting the whole tree together to make what can often be a time-sensitive decision. Thus to solve a finite or well-defined problem, it can be advantageous for AL1C3 to develop a fully detailed model for how best to proceed through an unknown environment. But in wide-ranging, multi-step situations, it quickly becomes impossible to

---

and rewards of a series of upcoming moves. But, just as in chess, where a novice player may have less accurate initial estimates than an expert player, so a model-based learner may have more and less accurate initial estimates of *R* and *T*. It is further possible that the habit-based system provides the goal-directed system with baseline estimates (Daw *et al.* 2011).

41

map all potential courses of action.

4.1.2. The Habit-Based (Model-Free) Decision-Making System

In contrast to goal-directed learning, as a habit-based learner, ALIC3 still aims to maximize its rewards, but it learns about its environment differently, namely, through trial and error. In order to do this, habit-based learning relies on a special signal known as an error prediction signal, which allows it to update its value calculations as it encounters grid states over time. In particular, as AL1C3 travels iteratively through a given grid space, it visits and revisits individual states more than once. For example, on its first mapping of the space, AL1C3 may encounter state B4 on its way to a dead end; but on its next four rounds through the grid space, AL1C3 may travel through B4 and find that it is also on the way to a valuable reward. This means that AL1C3 would value B4 relatively lowly on its first visit, but gradually upgrade its value over the course of its subsequent visits.

Notably, the process of re-valuing can also be described in terms of what AL1C3 'expected' to encounter in B4 and what reward it actually got when travelling through it the next time, or the next series of times. The error prediction signal enables an agent to match its expectations about what *will* happen against its experiences of what actually does happen. Recalling Redish's distinction between pleasure and reinforcement in Chapter 3, Section 5.2. (Figure 16 in previous chapter [all figures will be re-numbered to add up]), AL1C3 registers a positive windfall when it experiences a better than expected outcome, but experiences 'disappointment' when it anticipates a positive outcome that then fails to materialize.[42] The error-prediction system provides the agent with a highly effective and

[42] Steiner and Redish further distinguish between disappointment and 'regret.' Disappointment occurs when an agent anticipates a specific measure of reward or punishment, but it fails to materialize, so that she feels either disappointment or relief (depending on the positive or negative nature of what had been expected). By contrast,

accurate way of continually updating its knowledge about its environment.

One theory of how the error prediction signal is produced is known as the temporal difference (TD) algorithm, where the algorithm compares experienced reward against hitherto 'expected' reward (Barto *et al.* 1983; Sutton and Barto, 1998). One easy way to understand this relationship between expectation and outcome is to imagine an agent, Jim, going to a familiar French restaurant and ordering his favorite meal, *Boeuf Bourguignon*. Because Jim has eaten this meal many times before, he has certain expectations about how it will taste when he gets it.

Habit-based learning is also called 'model-free' learning and 'cache search.'

Box 2. More synonyms

When his meal arrives, he is able to see how this specific version of it measures up. On the one hand, he may say, 'This is the best plate of beef bourguignon I've ever had,' meaning that the meal is even better than he had come to expect it to be over time. On the other hand, Jim could say, 'The beef is a little tough for my taste, and the sauce is very salty today,' suggesting that this version of the dish didn't quite live up to his expectations. Or, finally, Jim could say, 'Chef Jacques has done it again,' indicating that the dish is exactly what he expected from it, and that it was tasty. Importantly, if Jim comes back several times and finds that the dish is consistently no longer as good as it used to be, he may learn from experience and stop ordering it. This would be an example of adjusting one's valuation over time, and correspondingly refining one's policy in an effort to optimize one's outcomes.

To generalize from the example, the TD algorithm enables an agent to weigh expectations against actual rewards, and to learn how to optimize actions over time. If the reward is larger than expected, the agent is able to gauge that things have gone 'better than

---

regret occurs when an agent recognizes that an alternative (counterfactual) action would have generated a more rewarding outcome (Steiner and Redish 2014).

expected'; if the reward turns out to be smaller than anticipated, the signal registers 'worse than expected.' If the anticipated and experienced rewards are the same, the signal records 'no change.'

More formally, the TD algorithm expresses this relationship between predictions and rewards. After each experience, an agent updates the value grid reflecting its environment to include its most recent experiences. As a result, its grid reflects the machine's current understanding of its surrounding environment. By weighing its expectations against the reality it subsequently experiences, it is able to make more accurate predictions over time.

Habit-based learning exhibits the opposite computational strengths and weakness of those of goal-directed learning. It does not apply the trial, modeling, and planning stages of goal-directed learning. Rather, it immediately enters the situation and updates its estimation of the best course of action as it proceeds. As a result, model-free learning is statistically inefficient, because it requires a very high number of experiences (samples) in order to arrive at a representative assessment of present and future values, i.e., to arrive at equilibrium where it is able to more or less correctly predict the value of a future experience.

In addition, habit-based learning is not immediately sensitive to unexpected changes. For example, if Jim has been eating Chef Jacques' beef bourguignon for years, he may be served a less tasty meal (i.e. the immediate reward will be low), but it will take two or even three bad meals before he reevaluates the whole experience sufficiently to stop ordering it altogether. Due to the 'summing' nature of the algorithm, it will take several low-reward experiences to lower the overall high value scoring that had heretofore been attributed to the beef bourguignon. Because habit-based learning relies on caching values

rather than making explicit predictions about them at each stage, an agent relying on habits typically fails to recognize when circumstances in its environment change. As Daw *et al.* (2005, 1705) observe, "working with cached values is computationally simple but comes at the cost of inflexibility: the values are divorced from the [immediate reward] outcomes themselves and so do not immediately change with the re-evaluation of the outcome." Together, the relative advantages and disadvantages of the goal-directed and habit-based mechanisms may explain why an efficient decision-making system would incorporate both systems, arriving at a kind of action-selection redundancy. The interactions between these two systems are discussed in Section 5 below.

4.2. Classical Conditioning and the Pavlovian Controller

Instrumental behaviors are not the only way agents can learn about and respond to their environments. Classical conditioning refers to cases where an agent learns that a certain stimulus predicts a positively or negatively significant outcome, but where this relationship is independent of any controlled response or behavior on the part of this agent. One key alternative is for a subset of behaviors to be 'hardwired' as part of an agent's repertoire, ensuring rapid, efficient, and reliable responses to certain types of stimuli. This third decision-making system is known as the Pavlovian controller and consists of a strict 'stimulus-response' relationship.

The name 'Pavlovian controller' can occasionally cause confusion. For this reason, it will be useful to specify the behavior and mechanism in more detail.

Classical conditioning is most frequently associated with Pavlov's original experiments with dogs, where Pavlov trained his dogs by repeatedly ringing a bell and then consistently feeding them afterwards. Famously, the dogs learned to associate the ringing

of the bell with the delivery of food, and the dogs' expectation of food delivery was measured by their salivation. In this context, the food was defined as the unconditioned stimulus, the salivation as the unconditioned response, and the bell as the conditioned stimulus. When the dogs salivated in response to the ringing of the bell, this was identified as the conditioned response.

In both public understanding and psychology, attention is paid to Pavlov's discovery of the bell as the conditioned stimulus and its association with the unconditioned response, the salivating. But for the purposes of reinforcement learning and decision-making, it is really the relationship between the unconditioned stimulus (i.e., the food) and the unconditioned response (i.e., the salivating) that is of interest (for an interesting discussion of how these two interpretive traditions view Pavlovian learning differently, see Rescorla 1988). Contrary to the context in which they were first studied, these behaviors are not learned, and they do not appear to be controlled by the animal at all (see Sheffield 1965). Rather, they are most likely the products of a lengthy evolutionary history, which has selected for a range of automatic, appropriate responses in the face of appetitive or aversive stimuli (Macintosh 1983). They mainly involve approach and withdrawal in association with appetitively or aversively valenced stimuli, respectively. While rewards tend to evoke approach, punishments appear particularly efficient at evoking behavioral inhibition.

In the context of reinforcement learning, Pavlovian control is additionally characterized by the fact that it persists despite being suboptimal (Bouton 2006). For example, pigeons will continue to peck at a switch even when food is withheld every time they do so (Williams and Williams 1969). Similarly, chickens are unable to withdraw from a food dispenser, even though approaching it is consistently associated with a withholding

of food (Herschberger 1986). These policy biases are generally appropriate in natural environments, but their lack of flexibility in experimental settings is revealing about the underlying mechanisms of control (Huys *et al.* 2012). It is this unconditioned type of 'Pavlovian' behavior that will play a central role in one of the categories of weakness of will.

Computationally, the Pavlovian system is a variation of the temporal-difference learning algorithm described in Section 4.1.2 above. Since Pavlovian learning is action-independent, the corresponding algorithm consists of the function of the reward expected from a given state, but without taking into account what action is taken. In slightly different terms, the algorithm is "a function measuring future rewards expected from a state, while ignoring actions… that is, strictly speaking, averaging out the agent's action choices as though they were just another source of randomness in state-state transitions" (Balleine *et al.*, 2009, 376-377). This being said, if an agent learns the values of available states, it can further use this knowledge to associate that value with events that, in its experience, lead to these states. This latter ability corresponds to the behavior of conditioned reinforcement, in which responses can be reinforced not just by the unconditioned stimulus, but also by the neutral, learned stimulus.

4.3 Summary

The reinforcement learning literature describes the computational underpinnings characterizing three main controllers: the goal-directed, habit-based, and Pavlovian controllers. Each mechanism implements a different computational strategy and correspondingly exhibits a different behavioral profile. For the most part, access to multiple decision-making systems should enable an agent to optimize and take advantage of each controller's respective strengths and weaknesses. As I suggest below, however,

overlapping controllers may also generate conflicting directives. The remainder of this chapter analyzes the nature of some of these conflicts and hypothesizes that they may generate what has typically been characterized as weakness of will.

5. Interactions Between Controllers and Kinds of Weakness of Will

Balleine *et al.* (2008), Dayan (2011), and others (Daw *et al.*, 2005) argue that an ideal decision-making agent would employ several complementary controllers. Each controller would use limited knowledge and exposure to the environment to calculate the most beneficial course of action, but would use a different computational strategy to do so. In particular, Daw *et al.* (2005, 1704) contend, "the difference in the accuracy profiles of different reinforcement learning methods both justifies the plurality of control and underpins arbitration. To make the best decisions, the brain should rely on a controller of each class in circumstances in which predictions tend to be most accurate." In this way, the benefits of multiple controllers include redundancy and different computational "sweet spots" (Dayan 2011).

There are, however, also costs associated with the running of these parallel systems. From an agent's perspective, multiple systems can occasionally result in competition and generate sub-optimal behavior. In addition, from an analytical perspective, the overlapping systems make it difficult for researchers to 'reverse engineer' the separate systems, thus making it difficult to understand the controller's respective functions and parameters.

In this section, I discuss some of the ways in which the three decision-making mechanisms compete. In particular, I argue that some of these interactions produce suboptimal behaviors corresponding to different categories of weakness of will. I identify

and discuss two categories of weakness of will in detail: a habit-based versus goal-directed category of weakness of will, and a Pavlovian versus goal-directed type of weakness of will. The latter category can further be subdivided into 'Pavlovian cognitive' and 'Pavlovian behavioral' types of weaknesses of will, where each of these can be elicited by a positively or negatively valenced stimulus.

5.1. Suboptimal Interactions between the Habit-Based and Goal-Directed Controllers

The normative, computational model of valuation predicts how the model-based and model-free controllers can interact efficiently. Using simulations, Daw et al. (2005, 1705) argue that their results "suggest that principles of sound, approximate, statistical reasoning may explain why organisms use multiple decision-making strategies and also provide a solution to the problem of arbitrating between them." In particular, they use Bayesian approximation methods to predict the circumstances under which each controller will dominate. They argue that the likely accuracy or inaccuracy in a given situation of each controller's prediction will serve as the deciding factor for 'choosing' between them.[43]

Here, inaccuracy (or 'uncertainty') is defined as an agent's ignorance of the true values located in its environment.[44] Daw et al. (2005) propose that each controller may be able to track its own relative uncertainty, or in other words, that it may be able to measure

[43] 'Accuracy' and 'choosing between them' are filler words here, representing a 'gray box' in the mechanism schema specifying that there must exist a function governing interactions between the two systems. Daw *et al.* (2005) specify a possible computational method for satisfying this function. However, it remains unclear whether a) this is actually the correct way to characterize this function or b) even if it is, *how* this representation would actually physically be carried out in the brain. See Chapter 5, Section 2 on Incompleteness below.

[44] Daw et al. (2005) are at pains to point out that ignorance is to be distinguished from risk. In the latter instance, the probabilities of reward may very well be stochastic, making them hard to predict, but they *can* be known.

the accuracy of its predictions about the environment over time.[45] On the basis of this, the authors propose that, "as in other cases of evidence reconciliation in neuroscience, such as multisensory integration, we suggest that arbitration between values is based on the uncertainty or expected inaccuracy of each" (2005, 1706). In this way, each controller uses past experience to 'estimate' how likely it is to be able to provide a good prediction about what to do. Next, these accuracy rates are compared. Finally, the winning controller calculates the values for the alternatives at hand and makes the subsequent decision between them. In this way, the probability of choosing an action for execution is directly proportional to its estimated value.

Vitally, due to their different computational strategies, model-based and model-free learning exhibit different 'uncertainty profiles.' There are two main sources of uncertainty in tree search. First, the agent must choose an appropriate model, on the basis of which it can then make decisions about the situation at hand. We can use the following analogy as an example. If Adela is playing chess and mapping out her final couple of moves, she may ask herself, 'does my current position on the board mean that I should focus on the King and the Rook, or that I should play for a draw?' She will then build up a decision tree according to which moves she thinks will be most relevant. At the same time, Adela may misread the situation and formulate a decision tree that doesn't accurately reflect her real options and/or their potential consequences. This 'model selection' step corresponds to one of the main sources of uncertainty in goal-directed learning.

Second, realistic situations often require trees so wide or deep that they require approximation techniques in order to be computable. These techniques include pruning

[45] Although traditional reinforcement learning models do not typically track retroactive accuracy self-measurements, the authors use Bayesian variations to model how such computations might work. For more detail, see Dearden et al. 1998, Mannor et al. 2004.

branches from the tree, exploring only a subset of paths, or exploring a certain path only partway. Daw et al. (2005) argue that these approximations generate computational "noise" that can result in a discernible amount of inaccuracy.

Habit-based search does not generate this kind of computational noise, because its computations are simple and recalled rather than reproduced at the point of making a decision. However, as in the example of Jim's gradual re-evaluation of the beef bourguignon, these iterative and recall-based aspects of habit-based learning are at the root of other kinds of uncertainty. In particular, habit-based learning is initially very ignorant about a new environment, and cannot respond to change quickly. It takes multiple iterative interactions with its environment to learn if a new harmful or beneficial state or event has arisen. This can lead the system to make mistakes while the controller takes the time to 'update' its knowledge about the grid.

Because of their opposite computational approaches and correspondingly different uncertainty profiles, Daw et al. (2005, 1704) interpret the two controllers as representing "opposite extremes in a trade off between the statistically efficient use of experience and computational tractability." On the basis of simulations, Daw et al. (2005) found that tree search is far superior early on in an environment, "because any moral of experience immediately propagates to influence the estimates of action values at all states; the effect of bootstrapping [or model-free caching] in dorsolateral striatal temporal difference learning is to delay such propagation, making the system less data- efficient" (2005, 1708). Nevertheless, there is some evidence that the model-based system can occasionally be overrun by the model-free system, even in instances where the former would be more ideal to use. I hypothesize that when an agent deploys the statistically more reliable, but at time T inaccurate, model-free valuation mechanism, it experiences weakness of will. Such a

scenario could unfold as follows.

As noted above, the goal-directed system is statistically more reliable in novel settings, where modeling potential outcomes enables an agent to predict the likely values of various outcomes. For example, it is very useful to plan one's route in a new, unfamiliar environment, when one doesn't have any reliable habits to fall back on. To do this, for example, one can use a map, or turn to someone else's description of the options available (Figure 4.6.).

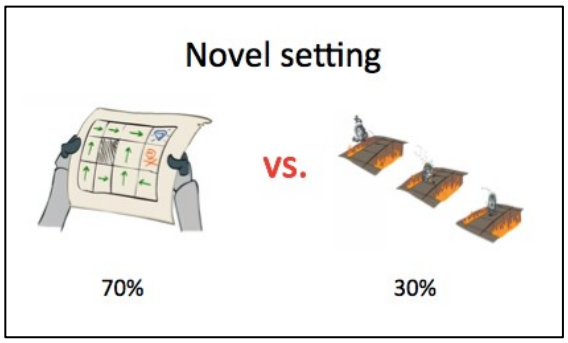

Figure 4.6. On the left, the goal-directed approach has a 70% probability of being accurate and efficient in the novel setting. On the right, the habit-based approach has only a 30% probability of being accurate and efficient. For example, provided I have a baseline of information available, I would probably map out my route in a new city rather than rely on habit (image modified from Abbeel & Klein, 2013).

By contrast, the habit-based system is much more efficient in complex but familiar circumstances, where tree-search would be both computationally taxing and redundant, such as taking a familiar route home (Figure 4.7).

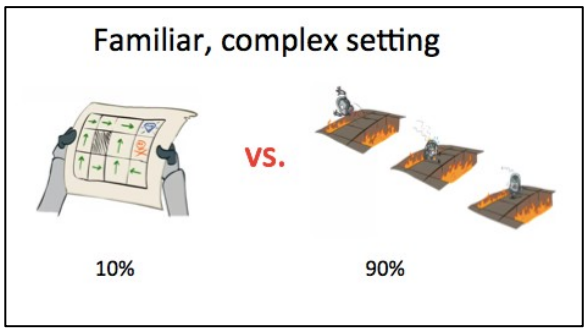

Figure 4.7. On the left, the goal-directed approach has a 10% probability of being both accurate and efficient in a familiar, complex setting. By contrast, on the right, the habit-based approach has a 90% probability of being accurate (image modified from Abbeel & Klein, 2013).

But it is not unusual for an important aspect of a familiar situation to change. For instance, one can rely on an established route home, but then have an accident or construction site appear unexpectedly. Together, these joint circumstances generate a situation where the habit-based prediction *appears* to be the more

accurate of the two, but isn't (Figure 4.8.).

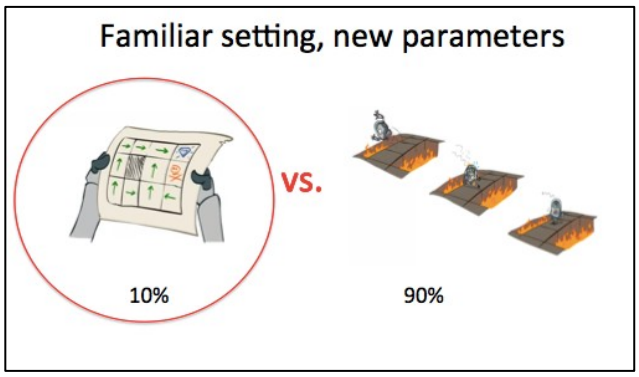

Figure 4.8. On the left, the goal-directed approach may be much less likely to be accurate than its habit-based counterpart, but it can nevertheless be the correct one under certain circumstances. This is especially likely to occur in familiar settings whose parameters have recently changed (image modified from Abbeel & Klein, 2013).

When an agent still opts for the more reliable (in past experience), but in fact inaccurate (in this new situation) habit-based controller, she will experience a moderate instance of weakness of will. The information provided by the goal-directed approach enables the agent to know what the best course of action would be under these recently changed circumstances. But since the situation is broadly familiar, the habit-based approach has a high past cumulative success rate, or an overall high accuracy rating, causing it to be 'selected' for the valuation task at hand. Correspondingly, the agent is aware of the most up-to-date and appropriate course of action, but 'falls back' on its less beneficial, habitual counterpart.

This characterization of weakness of will squares well with the fact that experiences of weakness of will frequently consist in actions that are well worn or 'habitual.' It would also suggest that this kind of weakness of will is remediable, because the habit-based system would catch on that the chosen course of action is disadvantageous. The habit-based model also accounts for Davidson's classic example of debating whether or not to brush his teeth. In what he calls a typical case of weakness of will, Davidson describes lying in bed at night and realizing that he's forgotten to brush his teeth. All things considered, he thinks to himself, it would be better just to stay in bed and

get a good night's sleep; but he just gets out of bed and goes to brush his teeth (1970, 30). As in the example of driving above, the habitual behavior 'wins out' over the modified circumstances of a familiar situation. Interestingly, although Davidson might get up once or twice to brush his teeth out of habit, he would very quickly learn that this is in fact a suboptimal course of action, and refrain from repeating it.

I propose to call this type of weakness of will 'Habit-Based weakness of will,' and suggest that it is more remediable and hence less detrimental than its Pavlovian counterpart, discussed below. Although it may not reflect the full-fledged 'struggle with temptation' implicit in some philosophical conceptions of weakness of will, e.g., St. Paul's account in his letter to the Romans, it captures the experience of an agent who knows she should not do something, but goes ahead and does it anyway. In this way, one major advantage of my multi-mechanism account of weakness of will is that it can accommodate the broader range of moments of weakness of will that individuals actually experience.

Redish himself describes just such an occasion during a visit to Tuscon to give a talk on multiple, interacting decision-making systems. He writes,

> I had come back to give a talk at the lab where I had done my postdoctoral work in Tuscon, Arizona. I was staying with my former mentors, in their guesthouse. When I worked in their lab (for about three years), I lived on the east side of town, on the south side of the airbase. They also lived on the east side of town, but north, by the mountains. So the first part of driving from the lab to their house was the same as the first part of what had been my usual trip home for years. And, *even fully conscious of this issue* that I had just laid out in my talk that morning, I turned south when I got to the corner and didn't realize my mistake until I had turned onto the street on which I used to live (2013, 43, Footnote A, added emphasis mine).

In other words, it was not that Redish did not know where his former mentors lived, or what the best way to get to their house was. Rather, it was that part of his route was so familiar that his habit-based model had a sufficiently high 'accuracy rating' to 'outweigh' the goal-directed system, and led him to drive home the regular – but now incorrect - way. As a result, Redish

likely brought his palm to his forehead, and experienced a moment of weakness of will.

5.2. Interactions between the Pavlovian and Goal-Directed Controllers

I hypothesize that a second category of weakness of will is elicited by suboptimal interactions between the Pavlovian and goal-directed controllers, and propose to call this Pavlovian weakness of will. I further suggest that Pavlovian weakness of will can be divided into two sub-types, which I will call 'Pavlovian Cognitive' and 'Pavlovian Behavioral' weaknesses of will. Pavlovian Cognitive weakness of will is characterized by a cognitive deficit hampering the decision-making process. Pavlovian Behavioral weakness of will refers to a Pavlovian approach or withdrawal behavior that hinders the agent from pursuing the most optimal course of action. Each of these sub-categories can involve either a strongly appetitive or aversive stimulus.

5.2.1. Pavlovian Cognitive Weakness of Will

In their paper, "Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees," Huys *et al.* (2012) examined the impact of the Pavlovian decision-making system on its goal-directed counterpart. In particular, they used a Bayesian model-fitting approach to see which decision-making system most closely corresponds to the responses participants provided on the decision-making tasks. This means they used statistical analyses to see how well the values predicted by each computational model actually 'fit' with the decision-making patterns they observed in their participants.[46] They found that in decision-making tasks, aversive stimuli caused

[46] To use an analogy, the process is something like matching symptoms to different potential diagnoses. If a patient has two run-of-the mill symptoms (e.g., coughing and a fever), a doctor might say that she is sick with one of several options and need to run more tests. But if one or more of the patient's symptoms are rather unusual, then the list of potential illnesses narrows and perhaps only one or two illnesses 'fit the bill.'

participants to prune their goal-directed or 'tree search' models. Let's look at this experiment in detail.

Huys and colleagues presented participants with a sequential decision- making task, where participants had to plan between 2 and 8 steps ahead in order to maximize their rewards. The participants were presented with a computer screen containing 6 state squares. At the start of each round, participants started out from one of these six squares, with the current state they were occupying highlighted in white. The participants' objective was to execute a sequence of moves to maximize their total reward, with the number of moves permitted in each episode marked in advance at the center of the screen.

From each state, participants could move to exactly two other states. The participants were asked to navigate using two random computer keyboard keys, 'U' and 'I,' which moved deterministically between the squares to the next two available states (Figure 4.9.). In order to learn the transition matrix governing the moves between squares, the participants underwent 40 practice trials. In each trial, they were placed in a random starting state and told to reach a random target state in up to 4 moves. Training continued until the participant reached the target in 9 out of 10 trials, at which point the participant was allowed to complete the main task. Each state square was also associated with a particular reinforcement and marked ('++,' '+,' '–,' and '––') to help the participants remember the reward or cost associated with each.

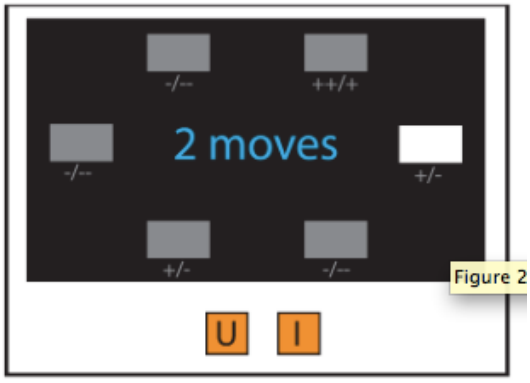

Figure 4.9. Sequential decision-making task from the perspective of the participants. For example, on this screen, the participants starts out from the white square, and must use the 'U' and 'I' keys to make two moves and accumulate the most reward possible (Figure from Huys *et al.* 2012).

The participants were divided into three experimental groups. The transitions costs were weighted differently for each group, so that each specific transition resulted in pre-determined amounts of reward or loss (Figure 4.10.). In particular, three of the six transitions cost the participants money. The first group's 'losing' transitions were weighted at -140. In practice, this meant that they earned the same amount of money over all, whether they pruned or not. By contrast, the second and third groups' losing' transitions were weighted at -100 and -70, respectively. In practice, these amounts caused the participants to earn less money overall when they pruned the decision-tree.

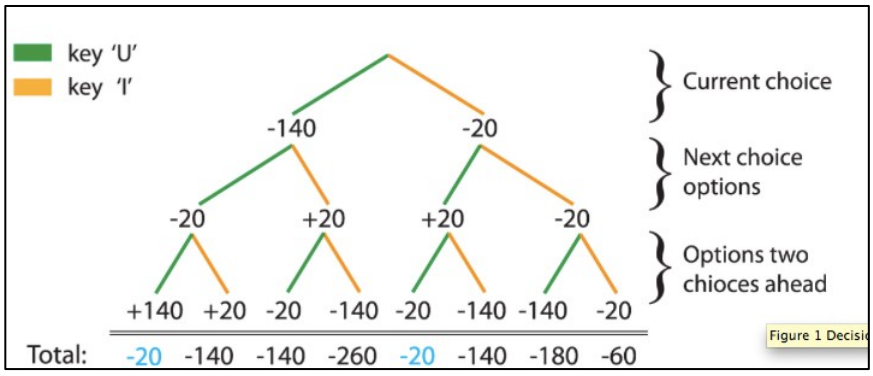

Figure 4.10. The 'back end' of the sequential decision-making task representing a possible sequence of choices. The large '-140' loss is of primary interest in this experiment, since it is predicted that it will cause participants to prune the branches in the decision-making tree (figure from Huys et al. 2012).

Huys and colleagues used a model-fitting approach to try and explain the participants' decision-making patterns they observed during the experiment. They

compared four models: 1) a 'look-ahead' model, wherein the participant would perform a complete tree evaluation, 2) a 'discount' model, where participants would only search the tree to a certain depth, 3) a 'pruning model', where participants limited their search by ignoring certain seemingly unprofitable branches, and finally, 4) a 'pruning and learned' model that supplement pruning (as characterized in model 3) with an additional Pavlovian component, intended to mimic attraction to experienced rewarding states and avoidance of states associated with punishment, or in other words, to incorporate past experience of positive and negative states into the decision making.

Huys *et al.* (2012) write that they developed this fourth model to mimic for "the conditioned attraction (or repulsion) to states that accrue with experience. This [model] captured an immediate attraction towards future states that, on average (but ignoring the remaining sequence length on a particular trial), were experienced as rewarding [in past trials]; and repulsion from states that were, on average, associated with more punishment." Based on their methods section, it appears that the authors initially expected the third, pruning-only model to accurately reflect the participants' decision-making processes. But over the course of the experiment, they found a pair of findings that they did not expect, and which the 'Pruning-only' model could not account for.

Specifically, the authors found that when participants had two moves left in a given episode, more than 90% of participants across all three groups made the optimal choice of moving through a small, -$0.20 loss (in the penultimate step) in order to then gain a large reward (in the last step). So far so good: this is what the 'pruning-only' model predicted. What it did not predict, however, and what surprised the authors, was that about 40% of participants also chose the -$0.20 loss on their *last* move! This means that they were willing to take a loss for no further gain, even when other, more

rewarding courses of action were available. This was obviously no longer the optimal choice to make. The authors hypothesized that some positive aspect of the previous trial, where the small loss had led to a large reward, influenced the subsequent, suboptimal decision. To try and account for this possibility, the authors developed the 'pruning and learned' model, which took into account participants' past experiences of rewards and loss, and how these experiences impacted their subsequent decision-making.

Of the four models, then, the added 'Pavlovian and Learned' model proved to be 'the best fit,' that is, it was best able to predict the participants' actual decision-making behavior. This suggests that individuals not only prune their cognitive options in response to strongly aversive or appetitive stimuli, but that they are also able to combined pruning with knowledge learned in previous trials – and that they do so when the earlier conditions no longer apply.

On the basis of this experiment, I propose that in Pavlovian Cognitive weakness of will, a vigorous negative or positive stimulus elicits a hard-wired Pavlovian response, resulting in the cognitive pruning of the model-based decision-tree. As a consequence, the agent's theoretical decision-making alternatives become restricted. In certain circumstances, this may cause her to feel as though 'she has no alternative' but to pursue a suboptimal course of action. For example, I propose that the Milgram obedience studies can be understood as an untapped set of instances of weakness of will, and that the mechanism of Pavlovian cognitive weakness of will may help explain the unusual behaviors observed in them.

The Milgram obedience studies are among the most famous experiments ever conducted in social psychology and feature an experiment – together with 18 variations –

designed to study obedience. In the baseline procedure, naïve participants were ordered to administer electric shocks to victims taking part in a 'learning experiment,' though, of course, unbeknownst to the participants, the shock generator was fictitious and the victims were confederates of the experimenters.

Although fourteen Yale seniors predicted that only 1.2% of participants would go through to the end of the shock series, 14 out of 40 participants refused to continue the experiment after some degree of shocking, while the remaining 26 obeyed the experimenters until the end. These participants punished the victims until they had reached the apparent 450 volts, the strongest shock possible, and the victim (who was out of sight) had already stopped responding (Milgram 1963).

Two features of the participants' behaviors, briefly mentioned by Maria Merritt, John Doris and Gilbert Harman's analysis of studies in their article entitled, "Character," in *The Moral Psychology Handbook* (2011), help substantiate my interpretation of the behavior exhibited in the experiment as corresponding to weakness of will. First, the participants generally demonstrated no desire to harm the victims of their own accord and, second, they exhibited extreme discomfort throughout the experiment, despite the fact that the majority went on through to the end of the session.

As Merritt *et al.* (2011, 364) highlight, the participants did not shock the victims of their own accord. Rather, over the course of the experiment, they were commanded to administer increasingly severe shocks, with the lead experimenter repeating a series of "prods" such as, 'Please continue,' 'The experiment requires that you continue,' 'It is absolutely essential that you continue,' and so on (Milgram 1974, 21-22). This suggests that the participants' behavior did not necessarily correspond to what they *wanted* to do.

Second, and even more strongly, even those participants who continued to administer shocks did so under extreme stress.[47] Milgram (1963, 6) describes some of symptoms of this extreme stress as follows:

> Many [of the participants] showed signs of nervousness in the experimental situation, and especially upon administering the more powerful shocks. In a large number of cases the degree of tension reached extremes that are rarely seen in sociopsychological laboratory studies. Subjects were observed to sweat, tremble, stutter, bite their lips, groan, and dig their fingernails into their flesh. These were characteristic rather than exceptional responses to the experiment.
>
> One sign of tension was the regular occurrence of nervous laughing fits. Fourteen of the 40 subjects showed definite signs of nervous laughter and smiling. The laughter seemed entirely out of place, even bizarre. Full-blown, uncontrollable seizures were observed for 3 subjects. On one occasion we observed a seizure so violently convulsive that it was necessary to call a halt to the experiment. The subject, a 46-year-old encyclopedia salesman, was seriously embarrassed by his untoward and uncontrollable behavior. In the post-experimental interviews subjects took pains to point out that they were not sadistic types, and that the laughter did not mean they enjoyed shocking the victim.

Although they are not often mentioned in cursory descriptions of the studies, these acute responses indicate that the participants' behavior towards the victims caused them to experience tremendous distress.[48]

In their analysis, Merritt *et al.* interpret these unique symptoms of distress as indicating that the participants do not endorse the violent punishment of the victim, but continue to press the button for other reasons. They describe these situations as ones where

---

[47] There were a few exceptions. Milgram does acknowledge, "After the maximum shocks had been delivered, and the experimenter called a halt to the proceedings, many obedient subjects heaved sighs of relief, mopped their brows, rubbed their fingers over their eyes, or nervously fumbled cigarettes. Some shook their heads, apparently in regret. Some subjects had remained calm throughout the experiment, and displayed only minimal signs of tension from beginning to end" (1963, 4).

[48] It could be argued that the participants' distress was caused by the suffering of the victim, regardless of their role in the situation; or that they were specifically distressed by their active role in causing the victim's suffering. Work done by Batson, Fultz, and Schoenrade (1987) suggests that it is the latter. In an effort to distinguish between distress and empathy, the authors defined personal distress as a "self-focused emotional response" involving "feeling alarmed, upset, disturbed, distressed, and/or perturbed" (1987, 21), and found that although witnessing the independently-caused suffering of other individuals does cause individuals to experience distress, it does not cause any of these extreme symptoms observed in the obedience studies. Witnessing harm not caused by a participant can provoke her to experience an increase in heart rate, distressed facial expressions, and self-reports of feeling anxiety or discomfort (Eisenberg *et al.* 1989). It does *not* cause her to laugh uncontrollably or experience seizures.

"individual subjects' behavior fails to comply with moral norms of the sort the subjects can be reasonably supposed to accept: the obligation not to inflict significant harm on an innocent person against his will, say, or the obligation to help others in an emergency if you are easily able to help, you don't have more important, conflicting obligations, and no one else will help if you don't. Since the subjects' demeanor often indicates that they (at some level) endorse the norms that their behavior contravenes, we suggest the label *moral dissociation* as shorthand for such phenomena" (2011, 363). Following a more traditional philosophical line of reasoning, I believe I am equally justified in recognizing these behaviors as instances of weakness of will. So how does the Pavlovian Cognitive mechanism schema explain this behavior?

The Pavlovian Cognitive model may explain this behavior insofar as the participants were faced with strongly aversive stimulus at the very top of their decision trees, namely, in the form of what must have felt like an immediate and negatively valenced confrontation with the 'experimenter' if they should refuse to administer the shock. Consequently, the prospect of this highly unpleasant interaction may have pruned their decision-making options, so that the 'stop shocking' course of action no longer seemed available. In the false binary of the experimental set up, this left only the alternative, antisocial course of action of shocking the fellow participant.

This interpretation is supported by findings from several of the experiment's variations. For example, in Experiment 14, the 'experimenter' is placed in the role of the learner, and a less authoritative figure called "Mr. March" instructs the naïve participant to shock him. Milgram writes, "Mr. March's instructions to shock the experimenter were totally disregarded… At the first protest of the shocked experimenter, every subject totally broke off, refusing to administer even a single shock beyond this point. There is no variation

whatsoever in response" (1974, 101-103). Along slightly different lines, in Experiment 15, the baseline experiment remains the same, but there are two 'experimenters' in the room with the participant instead of one. When the learner protests at the shock, the two experimenters verbally disagree with one another as to whether they should go on. In this version, 19 out of 20 participants did not continue administering the shocks past this point (Milgram 1974, 106). In both of these variations, the participant's decision-making options remain 'open': 'Mr. March' does not represent the same threat to the participant, while the disagreeing 'experimenters' seemingly keep the option of dissent on the table.[49]

I further hypothesize that a parallel pattern should also hold when an agent is faced with a strongly positive branch of a decision tree, she may correspondingly focus on that positive branch, at the expense of not exploring the less immediately rewarding branch (Fig. 4.11.).

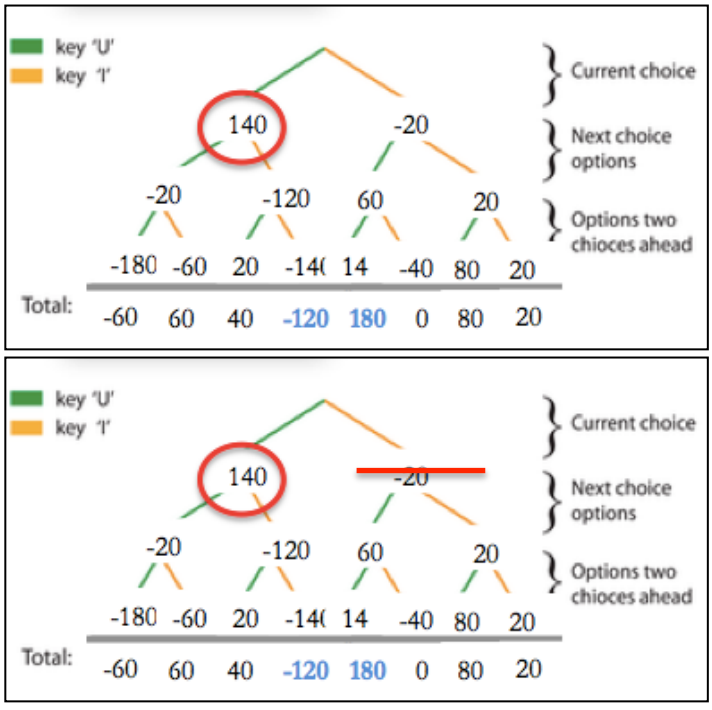

Figure 4.11. Top: The 'back end' of the sequential decision-making task representing a possible sequence of choices. This time, the large gain on the left is of primary interest, since it is predicted that it will cause participants to prune the branches in the decision-making tree. Bottom: The large gain causes the less immediately valuable branch to be pruned, despite the fact that this course of action proves to be less optimal overall (Figure adapted from Huys *et al.* 2012).

[49] In general, the majority of the variations suggest that the environmental circumstances that elicit profoundly anti-social human behavior are actually quite specific and, consequently, do not come together very often. (But when they do, they appear to elicit a 'perfect storm,' inhibiting almost all other-oriented behavior).

This positively valenced experience of weakness of will may help explain other traditional cases of weakness of will, including philosopher J.L. Austin's description of eating of a second slice of bombe. Recounting a positively valenced instance of weakness of will, Austin writes,

> I am very partial to ice cream, and a bombe is served divided into segments corresponding one to one with the persons at High Table: I am tempted to help myself to two segments and do so, thus succumbing to temptation and even conceivably (but why necessarily?) going against my principles. But do I lose control of myself? Do I raven, do I snatch the morsels from the dish and wolf them down, impervious to the consternation of my colleagues? Not a bit of it. We often succumb to temptation with calm and even with finesse (Austin 1956/7, 198).

Although it is the product of a 'hardwired' mechanism, one feature of Pavlovian cognitive weakness of will is that it needn't be rushed or impulsive. The pruning of the decision-making tree simply eliminates certain possible courses of action, potentially leaving the suboptimal alternative to be pursued "with calm and even with finesse" (Austin 1956/7, 198).

5.2.2. Pavlovian Behavioral Weakness of Will

I hypothesize that in Pavlovian Behavioral weakness of will, a vigorous stimulus elicits a hard-wired, Pavlovian response that compromises appropriate approach or withdrawal behaviors.

In a second paper entitled, "Disentangling the Roles of Approach, Activation and Valence in Instrumental and Pavlovian Responding," Huys *et al.* (2011) explored the relationship between the Pavlovian and goal-directed systems, this time focusing on the respective influences of valence (aversive vs. appetitive) and activation (withdrawal vs. approach behaviors). To do so, the authors set out from the well-established

psychological phenomenon of Pavlovian-Instrumental Transfer (PIT), in which an appetitive Pavlovian stimulus is known to enhance instrumental approach behavior.

For example, in PIT, a mouse is trained to associate a 'Hard Day's Night' with a food delivery. This is considered a Pavlovian association, because it is action independent: the mouse simply learns that, every time a 'Hard Day's Night' is played, a food delivery will follow, independently of what it is. Separately, the mouse is also trained to press a lever for food. This is considered an instrumental behavior, because the mouse must perform an action for a food delivery to follow (Figure 4.12.).

Name	**Definition**	**Example**
Pavlovian	Action independent	'Hard Day's Night'
Instrumental	Action dependent	Lever press
PIT	Pavlovian + Instrumental	Increased lever press

Figure 4.12. The central components of Pavlovian Instrumental Transfer

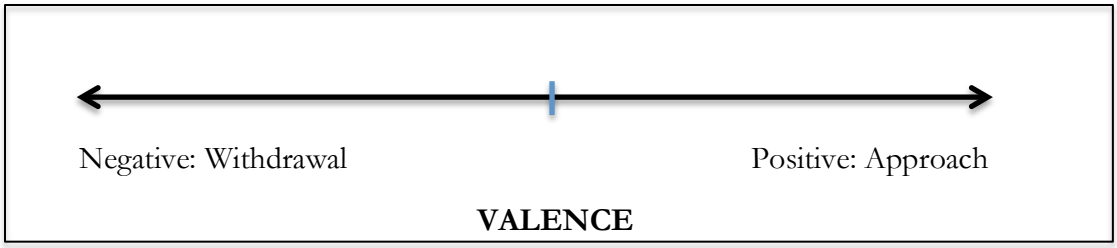

Figure 4.13. The 'Valence' axis. Left: Negative stimuli elicit Pavlovian (automatic) withdraw responses. Right: Positive stimulus elicit Pavlovian (automatic) approach responses.

PIT occurs when playing the famous Beatles' song causes the mouse to press the lever significantly more frequently than it would have done if no music were playing.[50] In this way, the Pavlovian stimulus appears to enhance a matching instrumental behavior (i.e., an instrumental behavior that shares the same 'approach' direction of behavior. In their research, Huys and colleagues were particularly concerned by the possibility that the shared 'approach' direction of both basic actions (i.e. the approach to the dispenser elicited by the song, and instrumental approach to the lever) was the driving force underlying the combined effect, namely, the overall, increased PIT approach.

Nevertheless, the authors felt that the 'enhancement' effect could not be clearly established, since both of the actions involved approach behaviors. They describe two motivational axes to help illustrate their understanding of the issue. The first axis corresponds to valence, namely, to positive or negative valence of different stimuli, and the corresponding direction of actions elicited (Figure 4.13.). This axis is pretty straightforward: as Socrates already noted, living organisms approach positively valenced stimuli, and avoid negatively valenced stimuli.

[50] In human beings, PIT is often invoked to explain factors contributing to addition. For example, if a former smoker walks past a restaurant where she used to smoke with her colleagues, she may feel an increased desire to go to a shop and purchase cigarettes, even if she hasn't smoked in years.

But things get tricky on the second axis. This axis corresponds to activation, where performing an instrumental action is typically tested in paradigms that pair 'complete task' with approach or 'go' behaviors, and 'failure to complete task' with staying put, or 'no-go' behaviors (Figure 4.14.).[51] This makes it difficult to tell whether the 'approach' behavior in the combined task is caused by the valenced Pavlovian stimulus or by the animal's instrumental attempt to complete the task in question. The authors observe, "the relative role of the appetitive-aversive motivation axis versus that of the approach-withdrawal axis is unknown. This in turn obscures the nature of the interaction: whether Pavlovian stimuli interact with the value of the instrumental behavior [i.e., whether it is the song that causes the mouse to act], or by promoting specific responses" that run 'in parallel' to the instrumental behaviors (i.e., whether the song and the instrumental task both cause the mouse to act) (Huys et al. 2011, 2).

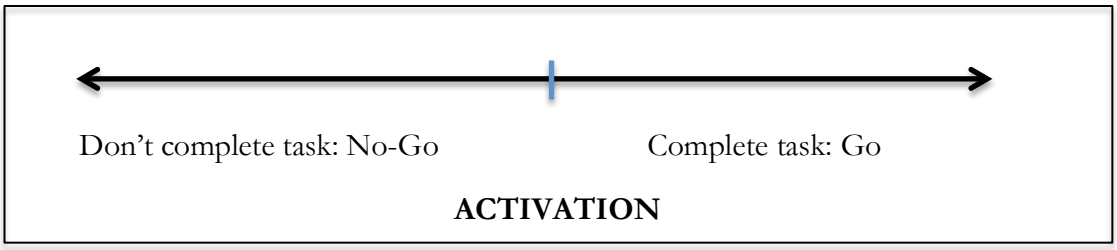

Figure 4.14. The 'Activation' axis. Left: Failure to complete the instrumental task typically involves 'no action,' or 'no-go,' e.g., sitting still and not approach the lever. Right: An attempt to complete the task typically involves action, or 'go,' e.g. going ahead and pressing the lever.

[51] Imagine the following problem. You want to investigate the effect of an incremental reward on a group of undergraduate students. As a researcher, you are very likely – in fact, you are almost guaranteed – to devise a paradigm which asks the students to *do* something, e.g. to press a series of buttons, in order to receive the rewards. So successful completion of the instrumental task involves a 'go' behavior, while failure to complete the task involves 'no go.' By contrast, it would be very atypical of you to ask participants *not* to do something in order to be rewarded. This would mean that successful completion of the instrumental task would involve 'no go,' and vice versa. Because of this tendency, in most experimental cases of PIT, the Pavlovian stimulus is positive, and so elicits approach, but so does the instrumental behavior, which typically asks for a 'go' behavior. The two thus create a confound.

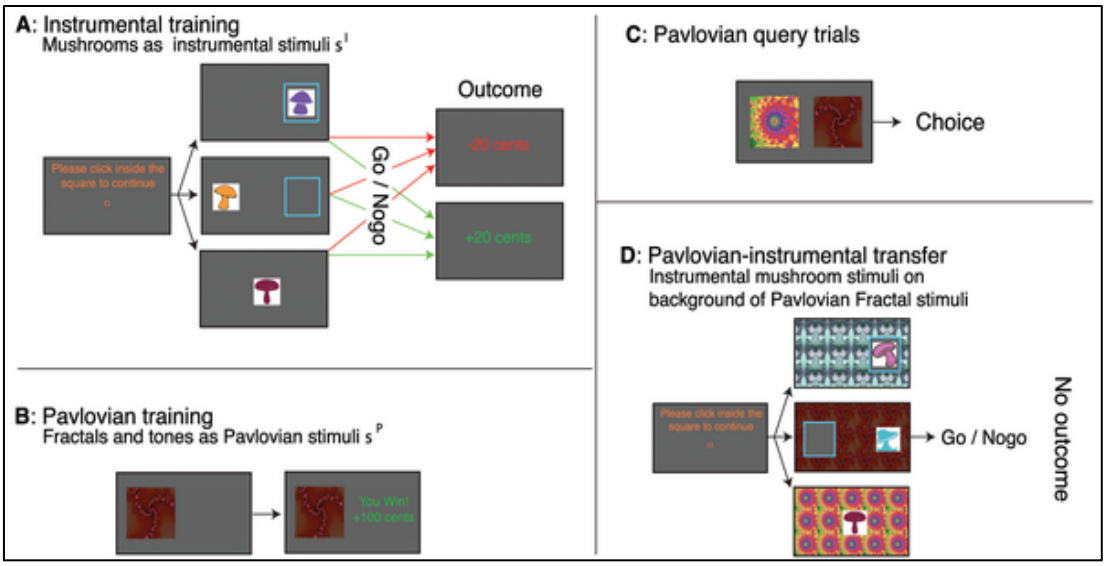

Figure 4.15. Description of the four tasks. A: The first tasks asked the participants to complete three types of instrumental tasks. (Top): The task with the purple mushroom asked participants to move the mushroom into the square (Go), or not to put into the square (No Go). (Middle): The task with the yellow mushroom asked the participants to drop the mushroom outside of the blue square (Go) or not (No Go). Bottom: The task with the red mushroom asked the participants to throw away the mushroom by releasing the button ('Go') or not throw away the mushroom by not releasing ('No-go'). B: The second task associated fractal tiles with positive and negative consequences. C: The third task tested to see whether the participants had learned the correct Pavlovian associations in (B). D: The fourth task examined the influence of the combined instrumental and Pavlovian tasks by asking participants to complete the instrumental tasks against the backdrop of the fractal tiles. (Figure from Huys *et al.* 2012).

In an effort to investigate the issue, the authors conducted an experiment exploring the relationship between valence and activation. In the study, the authors asked participants to complete four experimental tasks, involving both Pavlovian and instrumental settings, on a computer (Figure 4.15.). Regardless of whether they were in the Pavlovian or instrumental parts of the study, the participants were rewarded with + $0.20 and punished with - $0.20, with a running total kept throughout. The participants were allowed to keep the money they had earned at the end of the experiment.

The first task asked the participants to complete an instrumental task involving a mushroom on the computer screen. The task was framed in terms of mushroom collecting and sorting. There was one 'approach' version and two 'withdrawal' versions of the

instrumental task. In the 'approach' trials, the participants could choose to move the cursor to the mushroom and then to a blue square, thereby moving the mushroom into a blue square ('Go') or not do anything ('No-Go'). In the 'throw away' type of the withdrawal trials, the participants could choose to move the cursor away from the mushroom and then clicked in the empty blue frame to discard the mushroom ('Go') or not do anything ('No-Go'). Finally, in the 'release' version of the withdrawal trials, the mushroom appeared under the cursor at the beginning of the episode. The participants could choose to throw away the mushroom by releasing the button ('Go') or not throw away the mushroom by not releasing ('No-go') until 1.5 seconds had elapsed. After each episode, the participants were explicitly rewarded or punished by $0.20.

The second set of the experiment involved Pavlovian training. Here, the participants simply heard auditory tones and passively viewed stimuli consisting of fractal images. After each episode, the participants were explicitly rewarded or punished by $0.20, allowing them to learn the Pavlovian associations.

The third set of the experiment tested the Pavlovian association from the previous segment. Participants chose between two Pavlovian stimuli. The main difference here was that their gain and loss outcomes were not shown (although they were added to the participants' overall total at the end). This part of the experiment was designed to make sure the participants were still paying attention.

Finally, the fourth set of the experiment focused on Pavlovian-instrumental transfer. Here, the fractal Pavlovian stimuli from the second set tiled the background of the screen, and at the same time, the participants were asked to complete the same instrumental tasks from set 1. As in the previous set, the participants could not see their outcomes, but their choices were calculated in the final total.

In all, the participants were asked to complete tasks that both matched and counterbalanced valence with the direction of the instrumental task. In the *matched* cases, instrumental approach behaviors (e.g. move the mushroom into the blue square) were rewarded and paired with appetitive stimuli (e.g., the pink sunburst tile), while instrumental withdrawal behaviors (e.g., failure to move the mushroom into the blue square) were punished and paired with aversive_stimuli. (e.g., the maroon microbe tile). Alternately, again in the matched blocked, instrumental withdrawal behaviors were rewarded and paired with aversive stimuli, while instrumental approach behaviors were punished and paired with appetitive stimuli.

In the *counterbalanced* blocks, instrumental approach behaviors were rewarded but paired with aversive stimuli, while instrumental withdrawal behaviors were punished and paired with appetitive_stimuli. Or again, instrumental withdrawal behaviors were rewarded, but paired with appetitive stimuli, while instrumental approach behaviors were punished, but paired with aversive stimuli (Figure 4.16.).

Effect	Instrumental task	Reinforcement	Paired with
Matched	Approach	Reward	Appetitive stimulus
	Withdrawal	Punished	Aversive stimulus
Counterbalanced	Approach	Rewarded	Aversive stimulus
	Withdrawal	Punished	Appetitive stimulus
Matched	Withdrawal	Rewarded	Appetitive stimulus
	Approach	Punished	Aversive stimulus
Counterbalanced	Withdrawal	Rewarded	Aversive stimulus
	Approach	Punished	Appetitive stimulus

Figure 4.16. Fourth experimental set: matched and counterbalanced pairings of approach and withdrawal behaviors with appetitive and aversive stimuli.

The authors found that valence directly interacted with approach/withdraw behaviors in the PIT component of the experiment, even if withdrawal was a positively rewarded action. That is, a strongly appetitive stimulus (e.g., the pink sunburst tile) inhibited withdrawal behaviors, even on those blocks of the experiment where withdrawing was the positively rewarded instrumental action. Conversely, a negatively valenced Pavlovian stimulus (e.g., the maroon microbe tile) inhibited positively rewarded instrumental approach behaviors. These results indicate that the Pavlovian and goal-directed behaviors are dissociable, but can and do interact based on the 'direction' of the valence and activation.

On the basis of this experiment, I propose that Pavlovian Behavioral weakness of will occurs when a Pavlovian stimulus inhibits appropriate approach or withdrawal behavior. Further, the Pavlovian stimulus can be distinct and entirely unrelated to the 'approach or withdraw' task facing the participant. This suggests, for example, that an agent can intend to help a child in distress at the park, but be prevented from doing so because she encounters a particularly scary snake in her path.

When the valence of the Pavlovian stimulus overrides the most optimal course of goal-directed behavior, we can call this an instance of Pavlovian Behavioral weakness of will.

6. Comparing mechanisms

At first glance, it seems straightforward that my account of weakness of will has less in common with traditional, syllogism-based models of weakness of will than it does with its valuation-based counterparts. The three reinforcement learning models of decision-making do not share the structure of logical syllogisms, and, by extension, my account in no way

suggests that weakness of will is the product of a conflict between competing syllogisms. By contrast, my account does share the principles of valence, activation, and error. A distinct algorithm for valuation and corresponding behavioral profile defines each decision-making mechanism, and all three systems cohere with the approach and withdrawal behaviors fundamental to all living organisms. The multi-system account is also consistent with the view that, in an instance of weakness of will, an agent knowingly evaluates something as *apparently* more valuable in order to pursue it, even while this goal may nonetheless be *objectively* less valuable.

Despite these apparent similarities, however, three major differences distinguish my position from both historical models of weakness of will. First, my account rejects the view that there is a single system underlying practical reasoning, and instead follows the lead of reinforcement learning to hold the view that there is a three-mechanism model of decision-making. Second, my model breaks with the age-long view that weakness of will is a single phenomenon, and proposes that it is instead better understood as a cluster of behaviors. Third, I suggest that a key difference and advantage of my account lies in its ability to use Bayesian model-fitting techniques to determine the degree to which an agent is 'aware' of the detrimental consequences of her actions in a moment of weakness of will.

6.1. Inter-Systemic Vs. Intra-Systemic Competition

Traditional philosophical models of practical reasoning share in common the view that practical reasoning is underwritten by a single, unified mechanism. Many models discuss interactions between different mental faculties, including reasoning and the affects; perhaps most famously, Plato's discusses the tripartite soul in Part IV of the *Republic*. Nevertheless, these accounts describe intra-systemic competition between the faculties of a single,

centralized soul or mind, and not inter-systemic competition between different comprehensive systems. My account follows the lead of models from reinforcement learning to suggest that there are at least three complete, discrete decision-making mechanisms.

The move to multiple decision-making mechanisms is analogous to Dennett's efforts to decentralize conceptions of consciousness. In "Multiple Drafts Versus the Cartesian Theater," Dennett challenges the traditional notion of a unified consciousness and proposes a 'Multiple Drafts' model of decentralized processing. He writes, "all varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multi-track processes of interpretation and elaboration of sensory inputs" (1991, 111). He further suggests that conceptions of consciousness must shift from the traditional metaphor of writing up and editing to the ongoing editing processes made possible by word processing. He suggests that we

> consider a contemporary analogy. In the world of publishing there is a traditional and usually quite hard-edged distinction between pre-publication editing, and post-publication correction of "errata." In the academic world today, however, things have been speeded up by electronic communication. With the advent of word-processing and desktop publishing and electronic mail, it now often happens that several different drafts of an article are simultaneously in circulation, with the author readily making revisions in response to comments received by electronic mail. Fixing a moment of publication, and thus calling one of the drafts of an article the canonical text — the text of record, the one to cite in a bibliography — becomes a somewhat arbitrary matter" (1991, 125).

The transition from a single- to a multi-system model of decision-making requires an analogous shift in the way we understand the background conditions of practical reasoning in general and weakness of will of will in particular. In traditional models, an agent goes about her business, reasoning more or less rationally – until something 'goes wrong' and the

normal process is interrupted.[52] The multi-system model presents a different picture of the 'status quo' of our decision-making. We are what Montague (2011) calls "computational devices," with valuation systems continuously humming along in parallel within the mind, shaping our perceptions of and behaviors in the world. On this view, nothing 'breaks down' to produce instances of weakness of will. They are simply the byproducts of multiple systems optimizing our overall decision-making abilities.

6.2. Multiple Causes and Types of Weakness of Will

Second, my analysis is the first philosophical account to identify weakness of will as a *set* of behaviors, each with its own distinct computational mechanism and corresponding behavioral profiles. In doing so, it makes sense of some of the persistent disagreements that have surrounded weakness of will throughout the history of philosophy, including, 'Is weakness of will *a* thing?' and 'Why is there such a broad range of definitions of weakness of will?'

In Part 1, Chapter 1, I quoted the philosopher Stephen Schiffer's remark that weakness of will does even exist as a stable concept in the literature, but is rather "an unfortunate if picturesque term of art [that] has never had better than a vacillating reference" (1976, 201). I further proposed to bring some clarity to the issue by distinguishing between the behavior of weakness of will, or 'implementation failure,' and the many philosophical descriptions that have been offered for it; I suggested that everyday experiences of implementation failure are noncontroversial, and that it is only competing explanations of the phenomenon that have proved to be so thorny for professional philosophers. In light of

[52] And indeed, in Chapter 1, I suggested that this 'system breakdown' was the very research principle motivating investigations of weakness of will, with the *main* system of interest is actually the complex relationship between knowledge, motivation and action.

my multi-system analysis, I can now further argue that implementation failure is indeed non-controversial, but it is equally not a homogenous phenomenon.

The multi-system account helps explain why there is such a broad range of characterizations of weakness of will. In the *Protagoras*, Plato's Socrates' provides a very quotidian account of weakness of will that involves an agent simply doing something that she knows is not the best overall, such as eating too much or getting drunk. By contrast, philosopher Richard Dunn argues that only instances involving unconditional judgments about what is right constitute true, interesting moments of weakness of will. These definitions are much easier to accommodate when one has more than one type of weakness of will on hand. Several types of interactions produce a number of different behaviors, which collectively experience as 'weakness of will.'

6.3. Knowledge in Weakness of Will

The final issue continues with the question of knowledge in weakness of will. Does an agent really know that her actions are wrong in an instance of weakness of will? And if so, to what *degree* does she know that her actions are wrong? These questions are central to those who are interested in establishing standards of responsibility in weakness of will. Unfortunately, they are also notoriously difficult to establish, particularly when using an introspective approach to understanding weakness of will. The final and perhaps central advantage of my computation-based model of weakness of will is that it can use Bayesian model-fitting techniques to verify whether and to what degree the agent truly 'knows' the detrimental consequences of her actions in a given instance of weakness of will. By weighing participants' patterns of decision- making behavior against 'best fit' explanatory models, we are effectively able to see 'into' agents' minds to see what strategies are being

used to navigate their experimental environments.[53]

Consistent with my view that there is not one kind of weakness of will, my model shows that there is not a consistent degree of knowing involved in instances of suboptimal behavior. Indeed, the different models can specify the degree to which the agent 'knowingly' pursues a suboptimal course of action. In each case, establishing an agent's degree of knowledge will depend on the extent to which she was able to search her decision-tree before her action selection took place.

The uncertainty-weighing structure of Habit-Based weakness of will suggests that there may be some degree of parallel computation between the habit-based and goal-directed controllers. This in turns makes it very likely that an agent may be able to complete at least some portion of her tree search – and thus recognize the negative consequences of her choice of action – until the certainty associated with the habit-based system overrides it.

The case of the Pavlovian cognitive pruning is simultaneously more complicated and more interesting. At first glance, one could suggest that since the Pavlovian response immediately prunes one of the courses of action, the agent can never 'fully' know the consequences of her actions. On the basis of this, some philosophers would argue that this precludes this account from giving a meaningful account of weakness of will; in the sense that the agent has pruned the negative branch, or opted for the overwhelmingly positive branch, it does seem that a full tree search is no longer possible. Nevertheless, I think that the agent can, at a minimum, search down the non-pruned branch and see the consequences of her actions. And since the pruning causes her to select a less advantageous course of action, she pursues this action while being aware that it has

[53] In this way, 'best-fit' models can be far more valuable for resolving philosophical issues than topical fMRI experiments.

negative consequences. This combination of knowledge and behavior should, I think, meet the criteria of weakness of will.

Finally, it is likely that an agent can possess 'clear-eyed,' goal-directed knowledge of her circumstances in Pavlovian Behavioral weakness of will, since no pruning of the tree takes place. For example, it is quite likely that a participant in Huys et al.'s mushroom distribution task *knows* what the correct placement of the icon is, e.g., that she should move the mushroom into the box, but that her application of this knowledge is simply compromised by the opposing effect of the valenced stimulus in the background. In this way, the agent's knowledge of the situation is simply overridden by the parallel influence of the appetitive or aversive stimulus.

7. Conclusion

In this chapter, I have argued that weakness of will does not correspond to a single phenomenon, but actually consists in a cluster of behaviors, each with its own distinct computational mechanism and corresponding behavioral profiles (Figure 4.17.). I proposed that an agent experiences Habit-Based weakness of will when the brain deploys the statistically more reliable, but at time T inaccurate, habit-based decision-making controller, rather than its goal-directed counterpart. I suggested that an agent experiences Pavlovian weakness of will when a vigorous stimulus elicits a hard-wired, Pavlovian response. This can correspond to Pavlovian Cognitive weakness of will,

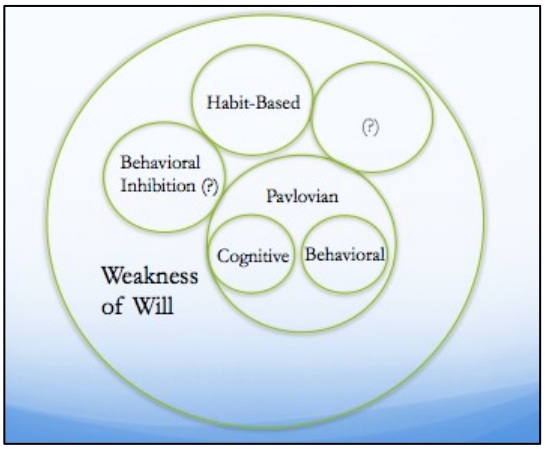

Figure 4.17. A cluster of weak-willed behaviors.

where cognitive pruning of decision tree limits an agent's theoretical decision-making alternatives. Alternatively, this can correspond to Pavlovian Behavioral weakness of will, where the Pavlovian stimulus inhibits appropriate approach or withdrawal behaviors. At the same time, I noted that the exact causal parameters distinguishing the Pavlovian Cognitive and Behavioral weakness of will remain in need of elucidation.

In Chapter 5, I return to Craver and Darden's criteria to assess the completeness and correctness of my multi-mechanism account of weakness of will. I will argue that, although it remains incomplete, my account has the potential to open up the remaining 'gray boxes' over time, and is thoroughly consistent with existing empirical evidence.

CHAPTER 5

BEHAVIORAL AND NEUROSCIENTIFIC EVIDENCE SUPPORTING THE HABIT-BASED AND PAVLOVIAN COGNITIVE AND BEHAVIORAL HYPOTHESES OF WEAKNESS OF WILL

1. Introduction

In Chapter 4, I developed a new, multi-system mechanism schema for understanding weakness of will. I examined the computational models characterizing the Pavlovian, goal-directed, and habit-based decision-making mechanisms. Based on my analysis, I argued that although weakness of will is traditionally identified as a single phenomenon, it in fact consists of a suite of discrete mechanisms that include 'habit-based' and 'Pavlovian' categories of weakness of will. In 'habit-based' weakness of will, an agent relies on familiar actions to navigate everyday situations, only to realize in some cases that the circumstances have changed and her actions are no longer appropriate. In 'Pavlovian' weakness of will, an agent's 'hard-wired' responses to threatening stimuli limit her ability to contemplate alternative courses of action; alternately, they inhibit her instrumental responses. I concluded the chapter by outlining three central differences between the traditional syllogism- and habit-based models of weakness of will and my own account.

Now it is time to ask, 'How well does my model really hold up?' In this chapter, I return to Craver and Darden's (2013) criteria for assessing mechanism schemas to evaluate my own account of weakness of will. I particularly focus on what they call the 'vices' of incompleteness and incorrectness. To address the first type of schema failure, in Section 2, I discuss two areas of my account that I believe correspond to explanatory 'gray boxes': namely, the as-yet underdetermined mechanism of 'arbitration' (Daw *et al.* 2005), as well as the probability of additional types of weakness of will. To assess the relative correctness or incorrectness of my account, in Section 3, I then turn to behavioral and neuroscientific

evidence suggesting that multiple decision making systems are indeed operational in animals and, especially, in human beings. In Sections 4 and 5, I propose ways to further test and substantiate my model of weakness of will. I argue that although it is incomplete, my mechanism schema for weakness of will is consistent with current findings from the psychology and neuroscience of decision-making.

2. Incompleteness and Types of Gray Boxes

Incomplete mechanism schemas can contain 'black' and 'gray' explanatory boxes. Black boxes represent components of the mechanism for which a function has not yet been identified. By contrast, Craver and Darden characterize gray boxes as representing functional sub-mechanisms that are believed to be involved but are not clearly understood. Not all gray boxes are oversights or turn out to be errors; they are simply underspecified. I propose to further distinguish the concept of gray boxes into 'essential' gray boxes and 'non-essential' gray boxes. The former correspond to functional sub-mechanisms that are necessary for a given mechanism to work. The latter correspond to functional sub-mechanisms that supplement or add to the mechanism.

In their analysis, Craver and Darden focus on what I call 'essential' gray boxes. They describe, for example, how Darwin recognized that his theory of evolution needed an account of the inheritance of variation, but was unable to provide it himself. They remark, "Darwin was fortunate that the problem of heredity could be cordoned off. For the purposes of Darwin's early theorizing, heritable variations could simply be documented to occur empirically (and Darwin had literally volumes of examples). At this stage of his discovery, he could afford to treat heredity as a form across generations produced by some-mechanism-we-know-not-what" (Craver and Darden 2013, 89). On this view, an essential

gray box lies at the heart of a given mechanism schema, and has the potential to 'make or break' how a mechanism works.

By contrast, a 'non-essential' gray box pertains to the as-yet underdetermined scope and boundaries of a given mechanism schema. For instance, many biological systems exhibit the characteristic of degeneracy, namely, the ability of different structural elements to perform the same function within a system (Tononi *et al.* 1999).[54] Instances of degeneracy occur at all levels of biological systems: from the level of genes, where different nucleotide sequences encode the same polypeptide, all the way up to the level of whole bodies, where different patterns of muscle contractions produce the same type of body movement (Edelman and Gally, 2001, 13764). Consequently, it is possible to understand some, without understanding all, of the mechanisms contributing to the generation of a specific phenomenon. In the absence of understanding all of these different contributing mechanisms, a researcher can formulate non-essential gray boxes to represent her hypotheses about further possible sub-mechanisms.

To help illustrate what I mean, let me return to the example of Hippocrates' study of dislocated shoulders from Chapter 1. Suppose that Hippocrates has identified three of the seven ways a shoulder can be dislocated, but has the sense that there are two other ways he just doesn't understand yet. His mechanism schema would then consist of three glass boxes and two gray boxes, but it would not represent quite the same type of incompleteness that characterizes Darwin's discussion of the inheritance of variation. Hippocrates' model could

[54] Although the two concepts are often confused and used interchangeably, degeneracy and redundancy are not the same. Redundancy "occurs when the same function is performed by identical elements," e.g., when two eyes perform the same function. They are redundant because if one eye gets poked out, the other can still see. Degeneracy, on the other hand, "involves structurally different elements, may yield the same or different functions depending on the context in which it is expressed" (Edelman and Gally 2001).

be quite accurate in the parts it has already identified, while also only capturing five out of the seven sub-components of the mechanism to varying degrees.

My multi-system model of weakness of will contains at least one essential and several non-essential gray boxes. The first, essential gray box corresponds to the mechanism of 'arbitration' between habit-based and goal-directed controllers (Daw *et al.* 2005).

2.1. The Essential Gray Box of Arbitration

In Chapter 4, I discussed how Daw *et al.* (2005) use Bayesian approximation methods to predict the circumstances under which the goal-directed and habit-based mechanisms each dominate. They argue that the likely accuracy or inaccuracy in a given situation of each controller's prediction will serve as the deciding factor for 'choosing' between them. In particular, they follow the lead of computational analyses of multisensory integration to specify a possible computational method for satisfying this function (Deneve and Pouget 2004). In their paper, Deneve and Pouget (2004) present a modified Bayesian framework to try and account for combinations of different sources of sensory information. They argue that multisensory integration is "a dialogue between sensory modalities rather that the convergence of all sensory information onto a supra-modal area," where the 'rules' of this dialogue are dictated by the estimates the reliability of different incoming sensory cues (2004). For example, Deneve and Pouget's model tries to account for the fact that for tracking the position of an object, visual cues are more reliable than auditory cues during the day, but that visual cues are also less reliable at night than during the day. Daw *et al.* propose to adopt similar principles of certainty and uncertainty to govern interactions between the habit-based and goal-directed mechanisms.

The concept of arbitration represents a central gray box in my mechanism schema. It

is a function that is essential to my understanding of habit-based weakness of will, but how the sub-mechanism actually works remains subject to debate. Notably, Daw *et al.* hypothesize that, "for simplicity, we assume that the estimated value of each action is taken to be that derived from the controller that is more certain about the value," and attempt to map out its computational underpinnings (2005, 1706 and supplementary methods). Taking up this hypothesis, researchers have attempted to elucidate the neural mechanisms of this arbitration process in rhesus macaques (Isoda & Hikosaka 2007) and humans (Lee *et al.* 2014). Lee *et al.*'s preliminary findings indicate that the human brain is indeed able to allocate degrees of control between model-based and model-free systems "as a function of the reliability of their respective predictions" (2014, 687). They additionally propose that the inferior lateral prefrontal and frontopolar cortices are responsible for encoding and comparing these respective reliability signals involved.[55] Nevertheless, these are only preliminary findings, and the issue has otherwise received very little computational or empirical attention. In this way, the task of elucidating the mechanism arbitrating between the goal-directed and habit-based systems seems to be 'on the right track,' but remains a debated 'gray' box rather than a full-fledged, established 'glass box.'

[55] Interestingly, Lee *et al.* also suggest that the model-base and model-free controllers may not share control equally over time. Rather, the efficient, model-free mechanism may represent the default system, which is only inhbited in circumstances that demand the taxing but precise model-based controller. They write, "It is notable that while we find evidence for effective connectivity between the inferior frontal and frontopolar arbitration regions and areas involved in model-free valuation in the putamen and supplementary motor cortex, we did not find any evidence for direct interactions between the arbitrator and regions involved in model-based valuation. These results imply an asymmetry in how the arbitrator operates: instead of modulating either model-based or model-free systems depending on which one has the most reliable estimate, the controller appears to work by selectively gating the model-free system. This could be consistent with the possibility that perhaps model-free control is in essence default behavior: unless the model-free controller has especially poor predictions, all else being equal (and due to reasons of computational efficiency), it is better for behavior to be under model-free, as opposed to model-based, control" (2014, 694).

2.2. Non-Essential Gray Boxes: Additional Types of Weakness of Will

The set of non-essential gray boxes in my mechanism schema corresponds to the probability that, apart from those I have already identified, additional types of weakness of will exist. I believe there are at least two promising avenues in the search for additional sub-mechanisms of weakness of will.

2.2.1. Tonic Immobility

First, what are known as 'freeze' behavioral responses may underlie a subset of moments of weakness of will. A 'freeze' response corresponds to something like temporary paralysis, or 'tonic immobility,' in the face of a threatening situation (Gallup 1974). For example, an animal may 'play dead' instead of fleeing from a predator. However, unlike the brief 'pause' sometimes observed in animals before they flee, 'freeze' responses appear to be automatic behaviors. As the psychologist David Barlow notes, "investigators have determined that tonic immobility is not a volitional […] on the part of the animal. Rather, this response represents another ancient behavioral reaction with obvious survival value. For the large number of predators for which attack is triggered and maintained by movement, freezing is an effective antidote that prevents further attack and increases the victim's chances of survival" (2002, 4). While relatively few studies have examined 'freeze' responses in human participants, it seems likely that humans similarly experience 'freeze' behaviors as non-volitional (Abrams *et al.* 2009), suggesting that freeze responses may 'feel like' doing something one clearly knows is not the best thing to do.In other words, they be similar to experiences of weakness of will.

In particular, in a study of tonic immobility in human beings, Schmidt *et al.* (2008) had participants inhale a combination of 20% CO2/80%O2 gas for 20 seconds. 20% of the participants described a strong desire to flee, and 13% of the participants described a feeling

of immobility. Surprisingly, however, the researchers found that that endorsement of the flight response, i.e., thinking, 'I should really flee!', was highly associated with the experience of the freeze response. This led the authors to suggest that, somewhat counter-intuitively, an "individual may experience immobility, combined with the wish to flee (but not [be] able to execute it)" (2008, 299). This experience corresponds to at least one characterization of weakness of will. Hare proposes that a moment of weakness amounts to an agent "sincerely assenting" to a command addressed to oneself, and *at the same time* not being able to carry it out due to a physical or psychological incapacitation (1963). Moreover, this tonic immobility does not appear to be an acute version of the behavioral inhibition elicited by the Pavlovian responses, as it exhibits both different behavioral (Maser *et al.* 1974, Sandberg *et al.* 1981) and neural (Monassi *et al.* 1997, 1999) profiles.[56] This suggests that in human beings, the 'freeze' response, or tonic immobility,' may represent a rare but acute type of weakness of will.

2.2.2. Vulnerabilities or Failure Modes in the Decision-Making Mechanisms

Second, Redish *et al.* 2008 (and Redish 2013) provide a different resource for navigating the search for additional sub-mechanisms of weakness of will. Their systematic analysis of intrinsic "failure modes or vulnerabilities" in the decision-making systems provides key clues for where to look next for sub-mechanisms and interactions between sub-mechanisms that may result in weakness of will.

To explain the concept of a failure mode or vulnerability, Redish (2013) discusses the opioid system and its three corresponding opioid receptors: mu [μ], kappa [κ], and delta [δ]. In particular, μ receptors appear to signal euphoria, and chemicals that stimulate them are generally

[56] For example, Maser *et al.* (1974) found that they were able to use Pavlovian responses to negative and positive stimuli to influence the probability and length of time that chickens (!) subsequently experienced tonic immobility. Negative stimuli increased both the likelihood and duration of a chicken's subsequent immobility; positive stimuli seemed to mitigate the effects.

euphorigenic. Herein, Redish notes, lies an important aspect of our physical constitution: euphoria is not an isolated mental experience, but rather a direct and inseparable product of a physical event in the brain. Moreover, it means that the right kind of chemical can produce the feeling of euphoria directly, and this makes the brain vulnerable to external influence. Redish writes, "We did not evolve μ-opioid receptors to take heroin; we evolved μ-opioid receptors so that we could recognize things in our lives that have value and thus give us pleasure. But heroin stimulates μ-opioid receptors directly and produces feelings of euphoria. Heroin accesses a potential failure mode of our brains," generating a feeling of euphoria that subsequently requires more and more heroin to recreate – until its user can no longer produce the feeling of euphoria at all (2013, 26). In this way, Redish argues, the μ-opioid receptors have a straightforwardly evolved function that simultaneously can make us vulnerable to profoundly suboptimal decision-making tendencies.

Taking the multi-part decision-making architecture into consideration, Redish *et al.* (2008) and Redish (2013) develop a taxonomy of vulnerabilities that target the goal-directed and habit-based systems and interactions between them. They include the overvaluation of expected outcomes, as when individuals remember past events as being much more pleasurable than they really were (e.g. binge drinking at a party), and the misclassification of situations, as when gamblers fail to distinguish between those situations in which they win money versus situations in which they lose money.

Redish *et al.* are at pains to point out that the taxonomy they provide is "certainly an incomplete list of the potential failure points of the decision-making system" (2008, 432). (And indeed, neither list includes the suboptimal interactions between the goal-directed and Pavlovian systems I propose in Chapter 4). In addition, it is important to note that Redish *et al.* (2008) mainly identify vulnerabilities that are exploited by drug use, and focus on systems that

result in severe dysfunction when compromised. For example, Redish *et al.* (2008) describe homeostasis and allostatic set points (i.e., permanent changes to an organism's homeostatic range) as susceptible to drug use,[57] as when repeated cocaine (Steiner & Gerfen 1998) and opiate (Cappendijk e*t al.* 1999) use result in permanent changes to the opiate receptors. These are certainly real exposures in the system; but these types of changes produce chronic decision-making patterns and behaviors that are far more severe than even the most counterintuitive cases of weakness of will (Gutkin and Ahmed 2011). They also appear to exhibit different neural profiles (Hyman and Malenka 2001, Kelley and Berridge 2002, Everitt and Robbins 2005).

This being said, Redish *et al.*'s list of vulnerabilities represent a good research option, because it represents a program interested in identifying 'weak links' in the decision-making architecture. Furthermore, it is likely that at least some of the suboptimal systems they describe may result in far less severe behavioral tendencies – including weakness of will – when not directly targeted by substances such as alcohol and drugs.

Investigating both of the gray boxes in my mechanism schema, namely, by examining the arbitration mechanism between goal-directed and habit-based learning, as well as by following up on Redish's taxonomy of vulnerabilities, should be undertaken alongside efforts to corroborate my existing hypotheses regarding weakness of will. As Craver and Darden note, the search for mechanisms is "frequently a piecemeal and protracted affair. It is piecemeal in the sense that one might work on a part of a mechanism, or an aspect of its function, while leaving much else about the mechanism inside a black box. It is also piecemeal in the sense that the different stages of discovery frequently interact with one another: one is forced to recharacterize

[57] Although Redish does not make this point, it is worth noting that allostatic set points can also be shifted by other factors including disease (Karatsoreo *et al.* 2011) and other types of chronic stressors (Juster *et al.* 2010), e.g., low social status (Howell *et al.* 2013)).

the mechanism in the face of learning about the mechanism, or one is forced to reevaluate experimental findings because one recognizes a previously unrecognized region of the space of possible mechanisms" (2013, 8). Consequently, incompleteness is best investigated alongside incorrectness to ensure that the discovery process is on course.

3. So What? Correctness and Incorrectness

A skeptic reading this dissertation may long have been wondering, 'So what? What's to say whether these decision-making mechanisms really exist? And how could we possibly ever tell them apart?' In this section, I argue that although my mechanism schema of weakness of will is not yet fully complete, it is correct, i.e., it is consistent with current scientific evidence. To use Craver and Darden's term, it is an 'how-actually schema' that "describes how the mechanism in fact works (or close enough for the purposes at hand." I support my claim by examining behavioral and neuroscientific evidence that suggests that multiple decision-making systems are indeed operational in animals and, especially, in human beings. Since there is less evidence regarding interactions between the systems, in Section 4, I go on to discuss how future research may further assess and corroborate my mechanism schema of weakness of will.

3.1. Behavioral Evidence

Researchers can distinguish between optimizing computational mechanisms based on their use of contrasting algorithms and their diverging accuracy profiles. It is much more difficult to identify and characterize decision-making mechanisms in living organisms. In an effort to pinpoint the nature of real-life decision-making mechanisms, researchers have devised numerous behavioral and neuroscientific assays for application in both animal and

human models. In this sub-section, I outline behavioral evidence supporting the existence of the three distinct Pavlovian, goal-directed and habit-based mechanisms.

3.1.1. Behavioral Evidence for the Pavlovian Decision-Making System

From a theoretical perspective, Pavlovian responses are thought to be the products of a lengthy evolutionary history, which has selected for a range of automatic, appropriate responses in the face of appetitive or aversive stimuli (Macintosh 1983). Perhaps not surprisingly, however, the non-instrumental aspect of the Pavlovian controller has frequently led to its characterization as being highly vulnerable to suboptimal outcomes, that is, by its tendency to persist even in those cases where it is detrimental to do so (Bouton 2006).

For example, in the mid-1960s, F.D. Sheffield devised an experiment to demonstrate that the dogs could not, in fact, control the salivation as a way to get more food. Adapted from Pavlov's paradigm, Sheffield's experiment involved a twist: if the dog salivated in response to the tone, it did not receive the food. This meant that the role of salivation would be shown to be either an action-independent response or as a deliberate, instrumental action geared towards receiving food from the experimenter. Specifically, Sheffield hypothesized that if salivation was a controlled action, then causing salivation to result in the withholding of food should result in the dogs not salivating.

Sheffield demonstrated that an action-independent stimulus-response relationship controlled the dogs' salivation. Even after over 800 tone–food pairings, the dogs salivated in response to the tone, even though this resulted in losing most of the food they could have obtained by withholding their salivary response. In this way, Sheffield showed that the dogs' salivation was not a deliberate action, and confirmed the existence of a separate

decision-making or action mechanism, which is known as the Pavlovian controller.

Using similar paradigms, Pavlovian responses have been elicited as a distinct behavior in other animals, including in pigeons and chickens. For example, pigeons will continue to peck at a switch even when food is withheld every time they do so (Williams and Williams 1969), and chickens continue to approach a food dispenser, even when doing so is consistently associated with the withholding of food (Herschberger 1986).

Pavlovian behaviors have also been studied with human participants. Redish (2013) suggests that somatic reactions, or automatic bodily reactions such as changes in heart rate and automatic facial expression, are among the most important Pavlovian responses (see also Damasio 2008). To illustrate his point, Redish offers a cheerful example from an experience working in his lab to illustrate his point. He writes,

> One time, when I was setting up my lab, I was worried about whether I could safely plug in a very expensive computer part without turning the computer off first […] I expressed my concern to the lab, and decided that I should be the one to plug it in so that if something went wrong, it would be my fault. While I was nervously waiting to plug the part in, unbeknownst to me, one of my grad students had snuck up behind me. The instant I plugged the part in, the student whispered 'bzzzzt!' in my ear. I must have jumped three feet in the air (2013, 66-67).

More formally, Pavlovian responses have been studied by looking at skin conductance responses, which can be used to look at experiences of fear, as well as at automatic facial expressions and emotional reactions.

Finally, Pavlovian interference may also play a role in behaviors such as impulsivity (Ainslie 2011) and other psychiatric diseases including anxiety and depression (Dayan and Huys 2008, Huys *et al.* 2011, Huys *et al.* 2012). For example, Huys *et al.* (2012) found that the higher the rates of cognitive pruning elicited by Pavlovian stimuli, the higher the rates of participants' own descriptions of experiencing depressive symptoms. This suggests that Pavlovian pruning may actually be related to the narrowed and more

negative worldview associated with depression.

3.1.2. Evidence Distinguishing the Habit-Based and Goal-Directed Controllers

Behavioral psychologists Dickinson and Balleine have used a technique known as 'post-training reinforcer devaluation' to differentiate between model-based and model-free behavior in animals. In reinforcer devaluation, researchers devalue a reward for the rats, either by pairing it with a nausea-inducing chemical or simply by overfeeding them with it, and examine whether the rats are still willing to press the lever to receive the pellet. Interestingly, the duration of the rats' initial training helps determine whether they are willing to press the lever or not: if they have only been trained for a moderate period of time, the rats appear to rely on model-based learning to conclude that since they don't want the pellet, they also don't want to press the lever. By contrast, rats that have been trained for longer periods of time appear to rely on the model-free learning system, and simply identify the lever pressing as a valuable action in itself. They continue to press the lever even long after they want to eat the pellets. In this way, both action sensitivity and outcome devaluation reveal key differences between model-based and model-free learning (Dickinson *et al.* 2002).

Work by Bitterman (1971) and Dickinson *et al.* (1998) demonstrates that some actions can also be controlled in direct proportion to how likely they are to achieve the desired outcome (see also Frankland *et al.* 2004). For example, Dickinson *et al.*'s study investigated rats' sensitivity to action contingency. During the training phase of the experiment, two groups of rats were trained to press two levers, A and B respectively, which delivered food pellets. The first group was trained for a short period of 4 sessions. The second group was trained for a longer period of 12 sessions. In the trial phase, both

groups of rats were subsequently introduced to a valuable sucrose solution, which was randomly presented while the rats were pressing levers for the food pellets. In the trial phase, pressing on lever A resulted in a withholding of the sucrose solution ('omission lever'), while pressing on lever B had no effect on the solution.

Dickinson and colleagues found that after the short training period, rats pressed lever A less and less often, suggesting that they still recognized the conditional relationship between lever A and the omission of the valuable reward. On the shorter training period, the rats continued to press the levers in a goal-directed manner. By contrast, the rats that had trained for 12 sessions no longer responded to the omission schedule. Rather, Dickinson and colleagues found that, after the longer training period, rats continued to press the A lever, even though doing so resulted in a withholding of the sucrose solution. This suggested that when the pressing of the lever became habitual, the rats were no longer able to manage the conditional relationship between the A lever and the omission of the valuable reward.

Tricomi et al. (2009) were able to distinguish between the habit-based and goal-directed systems in human participants using the same process of devaluation through satiation (Dickinson 1985). Tricomi and her colleagues fed one group of her participants M&Ms "until it was no longer pleasant to them" (2009, 4). She then observed that, if they were trained to press a lever for an M&M until the action became habituated, then the participants continued to press the lever, even if they no longer wanted the M&Ms.

To review, these experiments suggest that there are two distinct instrumental decision-making mechanisms, namely, the goal-directed and habit-based controllers.

3.2. Neural Evidence for the Existence of Pavlovian Values

Researchers have increasingly used lesion-based, recording-based (i.e., recordings of single neurons) and neuroimaging studies to uncover the neural underpinnings of the three decision-making controllers. In this section, I outline the emerging neuroscientific evidence supporting the existence of the distinct Pavlovian, goal-directed and habit- based mechanisms.

Neuroscientifically-oriented studies of Pavlovian values have focused on three key regions of the brain, namely, the amygdala, the orbitofrontal cortex, and the ventral striatum. The amygdala has particularly been implicated in stimulus-response behaviors in non-human primates. In a single-unit recording study undertaken by (Paton *et al.* 2006), researchers trained non-human primates to associate images with appetitive, aversive, or neutral values. For example, they taught a primate to associate a positively valued liquid reward with a picture of a pink sunburst and, conversely, they taught her to associate an aversive puff of air in the face with a maroon cell tile. Paton *et al.* then reversed the values assigned to all of the images, so that the pink sunburst was associated with the air puff, and the maroon cell tile was associated with the liquid reward. The researchers found that distinct amygdala neurons code for the positive and negative values associated with the images, and change rapidly enough when the associations are reversed to 'keep up' with the rates of monkeys' licking and blinking behaviors.[58] Additional studies have sought to work out details of the amygdala's role in valuation, including the timing and role of pre- existing

[58] Confusingly, the authors refer to the amygdala as providing the valuational basis for the monkeys' "*decisions* to either lick or blink during the performance of [the] task" (Paton *et al.* 2006, 5, added emphasis mine). However, licking and blinking are non-volitional 'stimulus-response' behaviors; they should not be called 'decisions.'

expectation in the process (Belova *et al.* 2007, Belova *et al.* 2008; but see also Murray 2007).

Anticipatory Pavlovian behavior has also been associated with monkeys' orbitofrontal cortices. Tremblay and Schultz (1999) found that reward processing in the orbitofrontal cortex is primarily related to relative preference of the available rewards, and not to any different physical properties. In particular, the authors found that activation in the selected neurons was related to the preference for the individual rewards; for example, in one trial, an apple was less preferable than a raisin, but in another trial, an apple was more preferable to a bowl of cereal. Tremblay and Schultz propose that neurons

> discriminate well between different rewards, and many discriminations appear to be based on the relative preference for different rewards exhibited by animals in overt choice behavior. The activity of these orbitofrontal neurons does not appear to code the fixed physical properties of rewards, but rather reflects the motivational value of one reward relative to another, as expressed by the behavioral preference. Just as each reward can have a higher or lower motivational value relative to the reward with which it is compared, orbitofrontal neurons can be more or less activated by one reward, depending on which alternative reward is available (1999).

This means that neurons on the orbitofrontal cortex appear to code directly for value and, further, for value in the context of the particular situation it is available in.

These preliminary findings have been also followed up in a wide range of studies focusing on the reward-coding properties of orbitofrontal neurons in non-human primates (Schultz *et al.* 1998, Schultz *et al.* 2000, Tremblay *et al.* 2000a, Tremblay *et al.* 2000b). The role of both the amygdala and the orbitofrontal cortex has equally been explored in human beings using fMRI, for example, by measuring hungry participants' responses to the smell of different tasty foods. The authors found that participants' responses to the smells stimulus decreased after they had been fed to satiation. By contrast, participants for whom the smell of food had not been devalued exhibited to same levels of activation throughout (Gottfried *et al.* 2003).

Finally, the ventral portion of the striatum has been implicated in Pavlovian valuation, a feature that squares well the finding that the ventral striatum is associated with the 'critic' architecture responsible for establishing reward expectations. For example, lesion studies in one part of the ventral striatum, the nucleus accumbens core, show impairment in Pavlovian approach behavior (Parkinson *et al.* 1999; see also Parkinson *et al.* 2000, Parkinson *et al.* 2002). In the 2000 study, Parkinson *et al.* found that removing part of rats' nucleus accumbens core disrupted the basic ability to pair positive stimuli with a corresponding approach behavior.

Together, these findings suggest evolved Pavlovian or 'state' values are encoded in a network of brain regions comprising the amygdala, orbitofrontal cortex, and ventral striatum (Balleine *et al.* 2009, 378). The question remains, however, how the brain is able to learn about and predict non-Pavlovian values in the external world. The answer to this puzzle lay in the breakthrough finding that something like a prediction-error signal (Chapter 4) exists and operates in the brain.

3.2.1. Prediction-error signal

The greatest discovery linking normative reinforcement models to actual biological processes was Wolfram Schultz and Peter Dayan's suggestion that the firing of dopamine neurons in relation to unexpected rewards amounts to a TD learning signal, exactly as predicted by optimized reinforcement learning. Read Montague, who helped work out the details of the correspondence, describes the discovery as follows. Contrary to the then-accepted view that dopamine reflects experiences of pure immediate reward, "Schultz noticed that dopamine neurons change their activity when 'important' events happened, like a juice squirt, or the appearance of food, or even a sound in the laboratory that

predicted that food or drink was about to be delivered" (2006, 108). This led him to the possibility that dopamine could correspond to a kind of prediction error. When he looked at Schultz's data, Dayan immediately "recognized a striking resemblance between dopamine neuron activity and error signals used in abstract reinforcement learning algorithms… it was an amazing match. The model showed that Schultz had discovered one of the central critic systems in the mammalian brain, and one that encoded its criticism in the delivery of dopamine" (2006, 109). In particular, just as the TD signal predicted, Schultz found that the dopamine neurons' spiking activity increased in response to an unexpected reward and decreased in response to an unexpected omission of reward. The main areas of activity implicated in Schultz's research were in the ventral tegmental area, which projects onto the ventral striatum, amygdala, and orbifrontal regions discussed above (Balleine *et al.* 2009).

Since the beginning of 2000, further research has demonstrated that something akin to a prediction-error signal operates in the brains of human beings. For example, Berns *et al.* (2001) highlighted the role of the nucleus accumbens and the orbitofrontal cortex. Relatedly, O'Doherty *et al.* (2003) found strong correlations between prediction error signal and the ventral striatum and orbitoprefrontal cortex, suggesting that the signal is present in human decision-making.

The neural correlates for the prediction-signal of aversive events is less well understood. Schultz was unable to record significant dopamine activity in relation to aversive events (1998)), though one study has even found that dopamine is inhibited during aversive experiences in rats (Ungless *et al.* 2004). In a theoretical paper, Daw *et al.* (2002) propose that serotonin may serve as a signal for predicting aversive experiences; however, work by Miyazaki, Miyazaki and Doya (2010) indicates that this is not the case.

These findings suggest that the prediction error signal is involved in the learning

and prediction of action states resulting in (at least) appetitive experiences, with neural correlates closely grouped around the ventral striatum. Together with the amygdala and orbitofrontal cortex, these regions of the brain appear to be strongly implicated in Pavlovian decision-making behavior, that is, in behaviors that do not involve the agent explicitly choosing an action in an effort to bring about a specific outcome.

3.2.2. Action Selection: the Actor/Critic Model

Finally, extensive research has been conducted to elucidate the neural structures of action selection. As noted in Chapter 4, instrumental decision-making involves at least two stages: a) learning about the reward outcomes of various action alternatives (carried out by the prediction error signal discussed in the previous section), and b) using the value predictions to select the next course of action. In habit-based learning, the first function is sometimes called the 'critic' and learns about different values based on experience. The second function is correspondingly known as the 'actor,' and together, they are known as the 'actor/critic' model of decision-making (Barto 1995). The question is, how are these two complementary functions realized in the brain?

In 1996, Read Montague and his colleagues proposed that the ventral striatum and dorsal striatum carry out the critic and actor functions, respectively. The critic system is clearly established in analyses of the prediction-error signal concentrated in the ventral striatum. In an fMRI study, O'Doherty *et al.* (2004) built on this finding to examine the role of the dorsal striatum in carrying out the corresponding (but separate) function of the actor. The authors reasoned as follows: "if the ventral striatum corresponds to the critic, then this region should show prediction error activity during both the instrumental and Pavlovian conditioning tasks. If the dorsal striatum corresponds to the actor, then we

would expect it to manifest stronger prediction error-related activity during instrumental than during Pavlovian conditioning" (O'Doherty *et al.* 2004, 452). And this is indeed what they found, suggesting that the dorsal striatum may play the complementary role of 'actor' to the function of 'critic' carried out in ventral striatum. Perhaps even more tellingly, lesioning the dorsal striatum in rats prevents them from being able to habituate behaviors (Yin *et al.* 2004; for a more detailed review of the valuational mechanisms in the striatum, see Knutston *et al.* 2009).

3.2.3. Neural Underpinnings of Goal-Directed Behavior

Finally, researchers using animal and human models have begun to identify the neural mechanisms associated with goal-directed behaviors. Working with rodents, for example, Balleine and Dickinson (1998) have demonstrated that the lesioning of either the prelimbic cortex or dorsomedial striatum results in rats being unable to execute goal-directed behaviors (see also Corbit and Balleine, 2003). More specifically, the prelimbic cortex appears to play a role in the preliminary learning of goal-directed behaviors, but not to be essential for carrying them out. Conversely, the dorsomedial striatum appears necessary for the both the learning and expression of goal-directed behaviors (Yin *et al.* 2005).

Working with human participants, Valentin *et al.* (2007) and her colleagues sought to explore which brain regions are active during goal-directed decision-making, that is, to examine which regions are active when the participants no longer sought out the devalued reward, indicating that they were able to respond to stimuli in a goal-directed way. Using fMRI, they found that participants who refrained from pressing a lever associated with receiving a devalued drink, i.e., chocolate milk, tomato juice or orange juice showed increased activation in their orbitofrontal cortices when making goal-directed decisions

(2007).

Using a slightly different paradigm, Hampton *et al.* (2006) used a decision-making task to try and pair a goal-directed decision-making pattern with activation in specific brain regions. To do so, Hampton and colleagues asked participants to complete a 'probabilistic reversal learning task.' In this kind of task, stimulus A and stimulus B are temporarily associated with a reward and loss, respectively, before the parameters of the association are reversed. As a result, throughout the trial, which stimulus is best to choose changes several times over the course of the experiment, and it is up to the participant to 'keep up' in order to gain as much reward as possible (in this case, the participants were rewarded with small sums of money, while a loss corresponded to having some of the money taken away following an incorrect choice). Hampton and colleagues found that when the participants were able to respond to the changes in the associations effectively, i.e., demonstrating goal-directed behavior, these choices were closely correlated with activation in the ventromedial prefrontal cortex.

4. Conclusion: Gathering Evidence for the Interactions Underlying Weakness of Will

In this chapter, I used Craver and Darden's criteria to assess the completeness and correctness of my multi-mechanism account of weakness of will. I argued that it remains incomplete, and discussed two 'gray boxes' that are in need of further clarification, namely, the functional sub-mechanism arbitrating between the habit-based and goal-directed controllers, and the possibility of additional types of weakness of will. At the same time, I suggested that my mechanism is thoroughly consistent with existing empirical evidence regarding the three decision-making systems.

Nevertheless, much more work remains to be done. The body of this chapter was devoted to laying out some of the evidence available in support of what I have described as the 'starting conditions' of my understanding of weakness of will, namely, the view that there are at least three distinct decision-making mechanisms. But these findings only indirectly support my hypotheses regarding weakness of will; they do not provide any direct evidence that *interactions* between these systems produce the kinds of weaknesses of will I hypothesize they do. By way of conclusion, I propose two future avenues for testing my explanations of weakness of will.

Recall that my three hypotheses propose that:

a) an agent experiences habit-based weakness of will when the brain deploys the statistically more reliable, but at time T inaccurate, habit-based decision-making controller, rather than its explicitly more appropriate goal-directed counterpart;

b) an agent experiences Pavlovian weakness of will when a vigorous stimulus elicits a hard-wired, Pavlovian response, where this can correspond to

    i) the cognitive over-pruning of branches of model-based decision tree, thus limiting her theoretical decision-making alternatives

    ii) the Pavlovian stimulus inhibiting appropriate approach or withdrawal behaviors

As a first line of investigation, I propose to use computer-based modeling to test both main hypotheses regarding the computational foundations of weakness of will.

Using the Neural Engineering Object (NENGO) software package (Eliasmith 2013), I propose to design two models to examine my mechanism schema. Consisting of different groups of neural structure and systems, the first model is intended to represent the neural circuits involved in the two complementary model-based and model-free decision-making mechanisms: the orbitofrontal cortex, the dorsolateral prefrontal cortex, dopaminergic neurons, serotonergic neurons, dorsolateral striatum, dorsomedial striatum, infralimbic

cortex, anterior cingulate cortex, and the basal ganglia. I propose to use this model to simulate interactions between the two systems, focusing on (i) whether and how these systems are normally integrated, (ii) whether and how the model-based system helps to train the model-free system (Dayan 2011) or vice versa (Daw et al. 2005), and, finally, (iii) whether and under what circumstances the model-free system 'overrides' the model-based system. Based on my analysis, I predict that this third set of simulations will generate results similar to those discernible in instances of weakness of will.

The second model is intended to represent the neural circuits involved in the Pavlovian and model-free decision-making mechanisms: the medial prefrontal cortex, the dorsomedial striatum, substantia nigra reticulate, mediodorsal thalamus, the orbitofrontal cortex, and the ventral striatum (Balleine et al. 2009). I propose to use this model to simulate whether and under what circumstances the Pavlovian controller 'overrides' the model-based system. Based on my analysis, I predict that this simulation will generate results similar to those discernible in instances of weakness of will, perhaps even more clearly and robustly than the first model can.

In addition, it will become essential in the future to work up direct, empirical support for my hypotheses regarding weakness of will. The next stage of research should involve adapting existing decision-making tasks with an eye to investigating the different types of weakness of will. A modified version of Hampton *et al.*'s (2006) reversal paradigm (discussed above) may be used to test habit-based weakness of will, and a version of Huys *et al.*'s (2012) pruning-based decision-making paradigm (discussed in Chapter 4) could be used to test Pavlovian cognitive weakness of will. If my hypotheses are correct, it should be possible to not only elicit different types of weakness of will, but also to work with participants to learn more about their subjective experiences of the phenomenon.

CHAPTER 6

CONCLUSION

In the preliminary stages of my dissertation research, an argument by philosopher Virginia Held caught and held my attention. In an article criticizing twentieth century philosophy's subservience to science, she reasoned,

> Cognitive science has rather little to offer ethics, and […] what it has should be subordinate to rather than determinative of the agenda of moral philosophy. Moral philosophers often make clear at the outset that moral philosophy should not see the scientific or other explanation of behavior and moral belief, or the prediction and control that science has aimed at, as our primary concerns. Our primary concern is not explanation but recommendation. I start from this position: ethics is normative rather than descriptive (1996, 70).

Held was by no means alone in arguing that cognitive science has nothing to contribute to ethics, but the strength of her formulation made it stand out. It was so categorical in its view that it was noticeably vulnerable to criticism. A single counterexample would be sufficient to refute it. 'Does such a counterexample exist,' I wondered? And so, after some thinking, I hit on the problem of weakness of will, or the phenomenon of acting against one's better judgment.

Ironically, I started out with a strictly casual – even complacent – attitude toward the problem of weakness of will.[59] All the same, this dissertation became a serious search for the mechanisms underlying it. My guiding principle was, 'Weakness of will occurs. Therefore, there must be a naturalistic explanation for it.' I moved beyond just wanting to prove Held wrong and became interested in how we are able to weigh and choose between different alternatives. I discovered that the mechanisms underlying valuation – the processes whereby we come to value and seek out what benefits us as living organisms, and avoid what is

[59]Which is to say, I didn't know anything about it. When I was asked to discuss Aristotle and Spinoza's contrasting views on akrasia on my comprehensive exams in September 2012, I passed over it because I didn't know what to say.

detrimental – could be within scientific reach. I began to want to really understand how these systems work, and I knew that if I could build on existing research, particularly from the computational neurosciences, I would be able to piece together an explanation for one aspect of decision making, weakness of will.

The first Part of this dissertation was devoted to navigating and organizing existing philosophical attempts to understand the mechanism of weakness of will. From the perspective of mechanisms, I suggested that there were historically two main schemas for doing so. I called them the 'syllogism-based' or 'valuation-based' models of weakness of will. The logical form of the syllogism provided the conceptual structure for the first of the two mechanisms. 'Valuation,' or the processes whereby we come to value and seek out what benefits us as living organisms and avoid what is detrimental, served as the guiding principle for the second of the two.

I proposed that syllogism-based models of weakness of will typically specify the following three principles:

Structure	That accounts of practical reasoning rely on deductive or inductive syllogisms
Conflict	That weakness of will arises out of a situation of conflict involving two syllogisms whose contradictory conclusions have opposite truth values and, finally,
Arbitration	That an affective faculty or an autonomous faculty of the will mediates between these competing syllogisms.

By contrast, I suggested that valuation-based models of weakness of will typically stipulate:

Valence	That agents attribute values to internal and external objects and events
Activation	That positively valuated objects and events elicit approach responses, while negatively valuated objects and events elicit withdrawal responses and, finally,

Error	That error is the product of agents evaluating an alternative as *apparently* more valuable than it actually is

The central aim of Chapter 2 was to show how syllogism-based models have systematically dominated philosophical treatments of the weakness of will throughout the history of philosophy, and how they remain prevalent even in contemporary philosophy (Harman *et al.* 2011). To this end, I analyzed the writings of two major philosophers who have advanced syllogism-based theories of weakness of will. In Section 3, I discussed Aristotle's account of weakness of will and showed how it became the prevailing philosophical treatment of the issue throughout much of medieval philosophy. In Section 4, I reconstructed Donald Davidson's analysis of weakness of will in his 1970 essay, "How is Weakness of the Will Possible?," and discussed Davidson's substantial influence on contemporary theories of weakness of will.

Chapter 3 outlined the alternate, valuation-based model of weakness of will. It analyzed the positions of Plato's Socrates (in both the *Protagoras* and *Republic*), Spinoza, and Hare and argued that, although they clearly set out from different philosophical points of departure, they shared in common the views that: a) an agent attributes values to internal and external objects and events, b) positively valuated objects and events elicit approach responses, while negatively valuated objects and events elicit withdrawal responses and, finally, c) although an agent must knowingly evaluate something as apparently more valuable in order to pursue it, this goal may nonetheless be objectively less valuable than the alternatives.

Equipped with a clearer sense of the history, in Part II of the dissertation, I presented my own mechanism schema of weakness of will. Broadly, I argued that converging evidence indicates that the human brain employs three dissociable mechanisms to make

choices. The 'Pavlovian' mechanism corresponds to 'hard-wired' approach and withdrawal responses. 'Goal-directed' behaviors map out different options and assess them in light of specific goals. 'Habit-based' behaviors learn the value of actions over time and in a given situation choose the most consistently valuable option in that situation. Although weakness of will is traditional identified as a single phenomenon, I argued that suboptimal interactions between these three decision-making mechanisms generate two different categories of weakness of will, which are etiologically but not psychologically distinguishable.

To make this clear, Chapter 4 considered how reinforcement learning in computer science investigates optimal decision-making systems, focusing on the computational models that characterize the Pavlovian, goal-directed, and habit-based decision-making mechanisms. I argued that weakness of will in fact consists of a suite of discrete behaviors that include 'habit-based' and 'Pavlovian' categories of weakness of will. In 'habit-based' weakness of will, agents rely on familiar actions to navigate everyday situations, only to realize in some cases that the circumstances have changed and their actions are no longer appropriate. In 'Pavlovian' weakness of will, agents' recognize the best course of action, but 'hard-wired' responses limit their ability to contemplate or pursue alternative courses of action.

Throughout the dissertation, I used Craver and Darden's criteria for evaluating mechanisms. I particularly focused on the 'vices' of incompleteness and incorrectness. Incomplete mechanism schemas are mainly characterized by the fact that they use placeholders where the relative sub-mechanisms involved in producing a phenomenon are not yet fully understood. A mechanism schema's correctness or incorrectness is evaluated based on whether one or more of its components are corroborated by, compatible or explicitly at odds with existing empirical evidence.

Based on these criteria, I argued that Aristotle, Davidson and 'Davidsonian'

accounts suffer from the 'vices' of incompleteness and incorrectness. I further suggested that these syllogism-based models of practical reasoning struggle to explain weakness of will because they lack a functional sub-mechanism responsible for evaluating better and worse alternatives. Similarly, I argued that the valuation-based models discussed underspecify key functional sub-mechanisms, but are somewhat compatible with current empirical evidence.

In Chapter 5, I used Craver and Darden's criteria for assessing mechanism schemas to evaluate my own proposal. I identified two areas of incompleteness in my account, namely, the as-yet underdetermined mechanism of 'arbitration' (Daw *et al.* 2005), and the unaddressed, additional sources of 'vulnerability' described in Redish's (2013) account of the three decision-making mechanisms. Based on behavioral and neuroscientific evidence suggesting that multiple decision-making systems are indeed operational in animals and, especially, in human beings, I argued that my account is correct, but needs further evidence to support its more specific hypotheses.

There are several features of this project that I would have liked to have done differently. For example, in Part I, I should have divided my historical analysis into three main mechanism schemas, i.e., syllogism-based, conflict-based, and valuation-based schemas, rather than to have too-forcefully tried to advocate for there being only the two syllogism-based and valuation-based accounts. Since many of my early ideas about how decision-making must work originated from the history of philosophy (particularly from the *Protagoras* and Spinoza's *Ethics*), I felt that I had not only found the seeds of my ideas in those texts, but that I must find the full-fledged ideas already present in these texts as well. This may have resulted in some oversimplifications that a historian of philosophy may rightfully find frustrating.

A scientist reading this dissertation may have the opposite problem of not finding enough empirical evidence to support my hypotheses regarding weakness of will. My mechanism schema is certainly mainly a theoretical-hypothetical proposal. I would have liked to develop full-fledged experimental proposals for how to investigate my ideas, but given my disciplinary limitations, I think not doing so may have been for the best for now. I certainly recognize the limitations of my account and, as I continue in my training, I hope to keep working on developing my ideas.

These and other issues notwithstanding, in searching for the details of the mechanisms of weakness of will, I see myself as having continued in the philosophical tradition of Plato's Socrates and those who followed in trying to understand the nature and implications of this strange phenomenon. I may have used some new resources, including some modeling from computational neuroscience, but even here I don't see myself as doing anything new. Many early modern philosophers also drew on the new mechanical sciences to try and explain this weakness in behavior, and a handful of contemporary philosophers continue to try to do so today (Levy 2011).

Craver and Darden argue that in uncovering mechanisms in the natural world, scientists proceed from 'mechanism sketches' to 'mechanism schemas.' The latter correspond to a "description of a mechanism, the entities, activities, and organizational features of which are known in sufficient detail that the placeholders in the schema can be filled in as needed," i.e. they may be said to have replaced any outstanding grey boxes with effective and verified glass boxes (2013, 31). I believe my explanation of weakness of will has the potential to become a 'glass box schema.' There are several instances where I use gray boxes, or 'filler terms,' to serve as placeholders for the statistical and neurochemical

interactions that take place. But I think I have my foot in the door, schema-wise, and I believe that further theoretical and experimental work will enable me to fill in the gaps.

If I have been able to make a contribution, then I have also made a small step toward responding to Virginia Held. Cognitive science has quite a lot to offer to ethics.

BIBLIOGRAPHY

Abbeel, P., & Klein, D. (2013). CS188 Artificial Intelligence, UC Berkeley [course notes]. Retrieved from http://ai.berkeley.edu

Abrams, M. P., Nicholas Carleton, R., Taylor, S., & Asmundson, G. J. (2009). Human tonic immobility: measurement and correlates. *Depression and anxiety*, *26*(6), 550-556.

Alanen, L. (2003). *Descartes's Concept of Mind*. Cambridge, MA: Harvard University Press.

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A.R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex, (Ed.) J. Cacioppo *Foundations in Social Neuroscience*. Boston: Massachusetts Institute of Technology, 2000, 333-344.

Anscombe, G. E. M. (1959/2000). *Intention*. Cambridge: Harvard University Press.

Aquinas, T. (1952). *The Summa theologica of Saint Thomas Aquinas,* Fathers of the English Dominican Province (Trans.). Chicago: Encyclopædia Britannica Press.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. London: Harper Collins.

Aristotle. (1984). *Nicomachean Ethics*, Bk. VII, Chs. 1-10., in *The Complete Work of Aristotle*, Ed. J. Barnes. Princeton: Princeton University Press, 1808-1821.

Audi, R., (1979). Weakness of Will and Practical Judgment. *Noûs* 13: 173-196.

Audi, R. (1990). Weakness of Will and Rational Action. *Australasian Journal of Philosophy* 68: 270-281.

Augustine. (1955). *Later works*, John Burnaby (Ed.). Westminster, John Knox Press.

Augustine. (1960). *Confessions*, John K. Ryan (Trans.). New York: Doubleday.

Arpaly, N. (2000). "On Acting Rationally Against One's Better Judgment," *Ethics* 110: 488-513.

Balleine, B.W. (2007). *Reward and decision making in corticobasal ganglia networks*. Boston: Blackwell Publishers.

Balleine, B. W., Delgado, M. R., & Hikosaka, O., (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, *27*(31), 8161-8165.

Balleine, B. W., Daw, N. D., & O'Doherty, J. P. (2009). Multiple forms of value learning and the function of dopamine. *Neuroeconomics: decision making and the brain*, 367-385.

Barlow, D. H. (2004). *Anxiety and its disorders: The nature and treatment of anxiety and panic*. Chicago: Guilford press.

Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. *Models of information processing in the basal ganglia*, 215.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *Systems, Man and Cybernetics, IEEE Transactions on*, (5), 834-846.

Baumeister, R.E., Bratslavsky, E., Muraven, M., and Tice, D.M. (1998). Ego Depletion: Is the Active Self a Limited Resource?. *Journal of Personality and Social* Psychology, Vol. 74, No. 5, 1252-1265.

Bechara, A., Damasio, H., Damasio, A. (2000). Emotion, decision-making and the orbitofrontal cortex. *Cerebral Cortex* 10 (3): 295-307.

Bechtel, W. (2002). Decomposing the mind-brain: A long-term pursuit. *Brain and Mind*, *3*(2), 229-242.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Philadelphia: Taylor & Francis.

Bechtel, W. (2011). Mechanism and Biological Explanation*. *Philosophy of Science*, *78*(4), 533-557.

Bechtel, W., & Richardson, R. C. (1993). Discovering complexity. *Princeton, NJ: Princeton UP*.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421-441.

Bechtel, W. and McCauley, R.N. (2001). "Explanatory Pluralism and Heuristic Identity Theory," in *Theory and Psychology*, vol. 11(6), 736-760.

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., & Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology*, 38(8), 1495-1503.

Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, *55*(6), 970-984.

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *The Journal of Neuroscience*, *28*(40), 10023-10030.

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy*, 49(188), 123-134.

Bennett, J. (1984). *A Study of Spinoza's Ethics*. Indianapolis, Hackett.

Bobonich, C. and Destrée, P. (eds.). (2007). *Akrasia in Greek philosophy: from Socrates to Plotinus*, Leiden: Brill.

Boureau, Y. L., & Dayan, P. (2010). Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, 36(1), 74-97.

Bouton, M.E. (2006) *Learning and Behavior: A Contemporary Synthesis*. USA: Sinauer.

Bratman, M. (1979(. Practical Reasoning and Weakness of the Will. *Noûs* 13: 153-171.

Brickhouse, T. C., & Smith, N. D. (Eds.). (1994). *Plato's Socrates*. Oxford University Press.

Brinckmann, P., Frobin, W., & Leivseth, G. (2002). *Musculoskeletal Biomechanics*. Stuttgart: Thieme Verlag.

Bruges, W. of. (1928). *Quaestiones disputatae du B. Gauthier de Bruges: texte* inedit. (Trans.) E. Longpre . Louvain: Institut supérieur de philosophie de l'Université.

Bunge, M. (2004). How does it work? The search for explanatory mechanisms. *Philosophy of the social sciences*, *34*(2), 182-210.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological review*, *58*(6), 413.

Buss, S. (1997). Weakness of Will. *Pacific Philosophical Quarterly* 78: 13-44.

Carone, G. (2001). Akrasia in the Republic: Does Plato Change his Mind?'. *Oxford Studies in Ancient Philosophy* 20, 107–148.

Casebeer, W.D. (2003). Moral Cognition and its neural constituents. *Nature Reviews Neuroscience* 4: 841-847.

Casebeer, W.D., Churchland, P.S. (2003). The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* 18: 169-194.

Charlton, W. (1988). *Weakness of Will*, Oxford: Basil Blackwell.

Churchland, P. (1998). Toward a Cognitive Neurobiology of Moral Virtues. *Topoi* 17: 83-96.

Churchland, P. (2007). *Neurophilosophy at Work*. Cambridge: Cambridge University Press.

Churchland, P.S. (1986). *Neurophilosophy: toward a unified science of the mind/brain*. Boston: MIT Press.

Churchland, P.S. (1995). Feeling Reasons, in A.R. Damasio, H. Damasio and Y. Christen (eds.), 1995, 181-200.

Clore, G.L., Huntsinger, J.R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Science* 11: 393-399.

Code, C. (Ed.). (1996). *Classic cases in neuropsychology* (Vol. 1). Psychology Press.

Cooper, J. M. (1984). Plato's theory of human motivation. *History of Philosophy Quarterly*, 3-21.

Cottingham, J. (1988). The intellect, the will, and the passions: Spinoza's critique of Descartes. *Journal of the History of Philosophy* 26:2, 239-257.

Craver, C., & Darden, L. (2001). Discovering mechanisms in neurobiology. *Theory and method in the neurosciences*, 112-137.

Craver, C., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago: University of Chicago.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, *329*(5987), 47-50.

Dahl, N. O. (1984). *Practical reason, Aristotle, and weakness of the will* (Vol. 4). U of Minnesota Press.

Damasio, A.R., Damasio, H., Christen, Y. (1995). *Neurobiology of Decision-making*. Berlin: Springer.

Davidson, D. (1970). How Is Weakness of the Will Possible?, in Davidson 1980, 21-42.

Davidson, D. (1978). Intending, in Davidson 1980, 83-102.

Davidson, D. (1980*). Essays on Actions and Events*. Oxford: Clarendon Press.

Darden, L. (1998). Anomaly-Driven Theory Redesign: Computational Philosophy of Science Experiments, in T.W. Bynum and J.H. Moor (Eds.), *The Digital Phoenix: How Computers are Changing Philosophy*. New York: Blackwell Publishers, 62-78.

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, *69*(S3), S354-S365.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4), 603-616.
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704-1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). 'Cortical substrates for exploratory decisions in humans.' *Nature*, *441*, 876-879.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current opinion in neurobiology*, *16*(2), 199-204.

Dayan, P. (2008). The role of value systems in decision making. In Engel C & Singer W, (Eds.), *Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions*. Frankfurt, Germany: MIT Press, 51-70.

Dayan, P., (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*(3), 213-219.

Dayan, P. (2011). Interactions Between Model-Free and Model-Based Reinforcement Learning. *Seminar Series from the Machine Learning Research Group*. University of Sheffield, Sheffield. Web. Apr.-May 2013. <http://ml.dcs.shef.ac.uk/>.

Dayan, P. (2011). Models of value and choice. In *Neuroscience of preference and choice: Cognitive and neural mechanisms*, ed. Raymond J. Dolan and Tali Sharot, 33–52. Waltham, MA: Academic Press.

Dayan, P. (2012). How to set the switches on this thing. *Current opinion in neurobiology*, 22(6), 1068-1074.

Dayan, P. (2012). Instrumental vigour in punishment and reward. *European Journal of Neuroscience*, 35(7), 1152-1168.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Dayan, P., Niv, Y., Seymour, B., & D Daw, N. (2006). The misbehavior of value and the discipline of the will. *Neural networks*, 19(8), 1153-1160.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2), 185-196.

Dayan, P., & Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience* 8 429-453.

Dayan, P., & Huys, Q. J. (2008). Serotonin, inhibition, and negative mood. *PLoS computational biology*, 4(2).

Dayan, P., & Huys, Q. J. (2009). Serotonin in affective control. *Annual review of neuroscience*, 32, 95-126.

Dayan, P., & Walton, M.E. (2012). A step-by-step guide to dopamine. *Biological Psychiatry* doi: 10.1016/j.biopsych.2012.03.008

Dayan, P., & Berridge, K.C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*.

De Houwer, J, Hermans, D. (Eds.). (2010). *Cognition and emotion: reviews of current research and theories*. New York: Psychology Press.

Della Rocca, M. (1996). Spinoza's Metaphysical Psychology. *The Cambridge Companion to Spinoza*, Ed. Don Garrett, 192-265.

Deneve, S., & Pouget A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology Paris* 98(1-3), 249-58.

Déscartes, R. (1985). *The Philosophical Writings of Descartes: Volume 1*, Cottinghman, J., Stoothoff, R., Murdoch, D., (trans.). Cambridge: Cambridge University Press.

Dickinson, A. (1985). Actions and habits: the development of a behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 308:67–78.

Dickinson , A., Squire, S., Varga, Z., and Smith, J.W. (1998). Omission learning after instrumental pre-training . *Q. J. Exp. Psychol.* 51, 271-286.

Dickinson, A. & Balleine, B. (2002.) The role of learning in motivation. In C.R. Gallistel (Ed.), *Learning, motivation and emotion*. New York: Wiley, 497-533.

Dihle, A. (1982). *The Theory of Will in Classical Antiquity*. Berkeley: University of California Press.

Doris, J. M., Cushman, F., and the Moral Psychology Research Group. (2012). *The Moral Psychology Handbook*. Oxford: Oxford University Press.

Doya, K., & Kimura, M. (2009). The basal ganglia and the encoding of value. *Neuroeconomics: Decision making and the brain*, 407-416.

Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest studies in philosophy*, *19*(1), 468-482.

Dunn, R. (1987). *The Possibility of Weakness of Will*. Indianapolis: Hackett.

Dworkin, R. (1996). Objectivity and Truth: You'd Better Believe It. *Philosophy and Public Affairs* 25(2):87-139.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763-13768.

Eliasmith, C., & Anderson, C. C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge: MIT Press.

Eliasmith, Chris. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature neuroscience*, *8*(11), 1481-1489.

Euripides. (2008). *Medea*, D. A. Svarlien (Trans.). Indianapolis, Hackett.

Fleming, P. (2010). Hume on Weakness of Will. *British Journal for the History of Philosophy*, *18*(4), 597-609.

Frijda, N.H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Geurts, D. E., Huys, Q. J., Den Ouden, H. E., & Cools, R. (2013). Aversive Pavlovian control of instrumental behavior in humans. *Journal of cognitive neuroscience*, 25(9), 1428-1441.

Ghent, Henry of (1979). *Quodlibet I*, R. Macken, ed. Leuven: Leuven University Press.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, 43(2), 127-171.

Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 443-464.

Glimcher, P. W., Fehr, E., Camerer, C., & Poldrack, R. A., Eds. (2008). *Neuroeconomics: Decision making and the brain*. Access Online via Elsevier.

Glüer, K. (2011). *Donald Davidson: A Short Introduction*. Oxford: Oxford University Press.

Gosling, J. (1990). *Weakness of the Will*. New York: Routledge.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, *301*(5636), 1104-1107.

Gutkin, B., & Ahmed, S. H. (2011). *Computational Neuroscience of Drug Addiction* (Vol. 10). Springer.

Grube, G. M. A. (1933). The Structural Unity of the Protagoras. *The Classical Quarterly*, *27*(3-4), 203-207.

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1), 154-166.

Gutkin, B., & Ahmed, S. H. (2011). *Computational Neuroscience of Drug Addiction* (Vol. 10). Springer.

Haidt, J., (2006). *The Happiness Hypothesis*. New York: Basic Books.

Hamilton, R., Hong, J., & Chernev, A. (2007). Perceptual focus effects in choice. *Journal of Consumer Research*, 34(2), 187-199. Chicago.

Hampshire, S. (1959). *Thought and Action* London: Chatto & Windus.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience*, 26(32), 8360-8367.

Hare, R. M. (1952). The Language of Morals, Oxford: Clarendon Press.

Hare, R. M. (1963). Freedom and Reason, Oxford: Clarendon Press.

Harman, G. (2002). Internal critique: A logic is not a theory of reasoning and a theory of reasoning is not a logic. *Studies in logic and practical reasoning*, *1*, 171-186.

Harman, G. (2011). Notes on Practical Reasoning." *Cogency* 3(4), 127-145.

Harman, G., Mason, K., and Sinnott-Armstrong, W. (2012). Moral Reasoning, in Doris *et. al*, 2012.

Held, V. (1996). "Whose Agenda? Ethics versus Cognitive Science," in *Mind and Morality: Essays on Ethics and Cognitive Sciences,* ed. Larry May, Marilyn Friedman, and Andy Clark, Cambridge: MIT Press, 70-87.

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, *14*(4), 443-451.

Hill, T. E. (2008). Kant on Weakness of Will. In Hoffman, T. (Ed.), *Weakness of Will from Plato the Present*, 210-230.

Hill, T. E. (2012). Kant on Weakness of Will. In *Virtue, rules, and justice: Kantian aspirations*, 107-128.

Hill, T. E. (2012). *Virtue, rules, and justice: Kantian aspirations*. Oxford: Oxford University Press.

Hippocrates. 1959. *On Joints.* (Translated by E. T. Withington). The Loeb Classical Library. Retrieved from https://archive.org/details/hippocrates03hippuoft.

Hoffmann, T. (Ed.), 2008, *Weakness of Will from Plato to the Present*. Washington: Catholic University of America Press.

Holton, R. (1999). Intention and Weakness of Will. *Journal of Philosophy* 96: 241-262.

Holton, R. (2003). "How is Strength of Will Possible?," in Stroud and Tappolet 2003, 39-67.

Howell, B. R., Godfrey, J., Gutman, D. A., Michopoulos, V., Zhang, X., Nair, G., & Sanchez, M. M. (2013). Social Subordination Stress and Serotonin Transporter Polymorphisms: Associations With Brain White Matter Tract Integrity and Behavior in Juvenile Female Macaques. *Cerebral Cortex*, bht187.

Huys, Q. J., Vogelstein, J., & Dayan, P. (2008). Psychiatry: Insights into depression through normative decision-making models. *Advances in neural information processing systems*, 729-736.

Huys, Q. J., & Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, 113(3), 314-328.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS computational biology*, *7*(4).

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, 24(6), 544-551.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3).

Hyman, S. E. (2007). The neurobiology of addiction: implications for voluntary control of behavior. *The American Journal of Bioethics*, *7*(1), 8-11.
Hyman, S. E., & Malenka, R. C. (2001). Addiction and the brain: the neurobiology of compulsion and its persistence. *Nature reviews neuroscience*, *2*(10), 695-703.

Irwin, T. (Ed.). (1995). *Classical Philosophy: Plato's ethics* (Vol. 3). Taylor & Francis.

Isoda, M., & Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature neuroscience*, *10*(2), 240-248.

Juster, R. P., McEwen, B. S., & Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, *35*(1), 2-16.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143-157.

Kalis, A., Mojzisch, A., Schweizer, T. S., & Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 402-417.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. *Punishment and aversive behavior*, 279-296.

Karatsoreos, I. N., & McEwen, B. S. (2011). Psychobiological allostasis: resistance, resilience and vulnerability. *Trends in cognitive sciences*, *15*(12), 576-584.

Kent, B. (1989). 'Transitory Vice: Thomas Aquinas on Incontinence,' *Journal of the History of Philosophy*, 27 (2), 199-223.

Kent, B. (1995). *Virtues of the will: the transformation of ethics in the late thirteenth century*. Washington, D.C.: Catholic University of America Press.

Klosko, G. (1980). On the Analysis of" Protagoras" 351B-360E. *Phoenix*, 307-322.

Knutson, B., Delgado, M. R., & Phillips, P. E. (2008). Representation of subjective value in the striatum. *Neuroeconomics: Decision making and the brain*, 389-406.

Korsgaard, C., 2006, "Morality and the Distinctiveness of Human Action,"in Frans de Waal, *Primates and Philosophers*. Princeton: Princeton University Press.

Kuhn, T. S. (1962/2012). *The structure of scientific revolutions*. University of Chicago press.

Lauwereyns, J. (2010). *The Anatomy of Bias: How Neural Circuits Weigh the Options*. Cambridge, MA: The MIT Press.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, *81*(3), 687-699.

Leibniz, G.W.F. (1965). *Nouveaux Essais Sur L'Entendement Humain*, Hans Heinz Holz (trans. German). Darmstadt : Wissenschaftliche Buchgesellschaft.

Lesses, G. (1987). Weakness, reason, and the divided soul in Plato's Republic. *History of Philosophy Quarterly*, 147-161.

Levy, N. (2011). Resisting Weakness of the Will. *Philosophy and Phenomenological Research* 82 (1): 134-155.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Science*. 8, 529–566

Lin, M. (2006). Spinoza's Account of Akrasia. *Journal of the History of Philosophy*, 44:3, 294-414.

Litt, A., Eliasmith, C., & Thagard, P. (2008). Neural affective decision theory: Choices, brains, and emotions. *Cognitive Systems Research*, *9*(4), 252-273.

Lloyd, G. (1990). Spinoza on the Distinction between Intellect and Will. E. M Curley and Pierre-François Moreau (Eds.), *Spinoza: Issues and Directions: the Proceedings of the Chicago Spinoza Conference*, Leiden: Brill.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, 1-25.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

Magnani, L., Nersessian, N., & Thagard, P. (Eds.). (1999). *Model-based reasoning in scientific discovery*. Springer.

Maser, J. D., Gallup, G. G., & Barnhill, R. (1973). Conditioned inhibition and tonic immobility: Stimulus control of an innate fear response in the chicken. *Journal of comparative and physiological psychology*, *83*(1), 128.

McCauley, R. N. (1999). Levels of explanation and cognitive architectures. *The Blackwell Companion to Cognitive Science*. Bechtel, W. and Graham, G. (eds.). Oxford: Blackwell Publishers, 611-624.

McCauley, R. (2012). About face: philosophical naturalism, the heuristic identity theory, and recent findings about prosopagnosia. *New perspectives on type identity: The mental and the physical*, 186.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339-346.

McClure, S., Botvinick, M., Yeung, N., Greene, J., Cohen, J. (2007). Conflict Monitoring in Cognitive-Emotion Competition, in J. Gross (Ed.), *Handbook of Emotion Regulation*. New York: Guilford, 204-228.

McIntyre, A. (1990). Is Akratic Action Always Irrational?. *Identity, Character, and Morality*. O. Flanagan and A. Rorty (eds.), Cambridge, MA: MIT Press, pp. 379-400.

McNeil, B.J., Parker, S.G., Sox, H.C., Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306, 1259-1262.

Meehan, W. (2009). Partem Totius Naturae Esse: Spinoza's Alternative to the Mutual Incomprehension of Physicalism and Mentalism in Psychology," in *Journal of Theoretical and Philosophical Psychology*, 29(1), 47- 59.

Melden, A.I. (1961). *Free Action*. London: Routledge.

Mele, A. (2010). Weakness of will and akrasia. Philosophical Studies, 150(3), 391–404.

Mele, A. (2012). *Backsliding: Understanding Weakness of Will*. Oxford: Oxford University Press.

Merritt, M. W., Doris, J.M., & Harman, G. (2011). Character. *The Moral Psychology Handbook*, Ed. John M. Doris. Oxford: Oxford University Press.

Milgram, S., 1963, Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4), 371.

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. London: Tavistock.

Miller Jr, F. D. (1999). Plato on the Parts of the Soul. *Plato and Platonism*, 84-101.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. Cambridge, MIT press.

Monassi, C. R., Hoffmann, A., & Menescal-de-Oliveira, L. (1997). Involvement of the cholinergic system and periaqueductal gray matter in the modulation of tonic immobility in the guinea pig. *Physiology & behavior*, *62*(1), 53-59.
Monassi, C. R., Ramos Andrade Leite-Panissi, C., & Menescal-de-Oliveira, L. (1999). Ventrolateral periaqueductal gray matter and the control of tonic immobility. *Brain research bulletin*, *50*(3), 201-208.

Montague, R. (2006). *Why Choose This Book? How We Make Decisions*. New York: Dutton.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*(6551), 725-728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience*, *16*(5), 1936-1947.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265-284.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760-767.

Montague P.R., Dolan R.J., Friston, K.J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Science* doi:10.1016/j.tics.2011.11.018

Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in cognitive sciences*, *11*(11), 489-497.

Nagel, T. (1978). Ethics as an Autonomous Theoretical Subject. In G. Stent

(Ed.) *Morality as a Biological Phenomenon*. Berkeley: University of California Press. 198-208.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in cognitive sciences*, 10(8), 375-381.

Niv, Y., & Montague, P. R. (2008). Theoretical and empirical studies of learning. *Neuroeconomics: Decision making and the brain*, 329-50.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current opinion in neurobiology*,*14*(6), 769-776.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329-337.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-454.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157-166.
Ovid. (2009). Metamorphoses, A.D. Melville (Trans.). Oxford: Oxford University Press.

Ong-Van-Cung, K.S. (2003). Indifférence et irrationalité chez Descartes. *Dialogue: Canadian Philosophical Review,* 42(04), 725-74.

Ostlund, S. B., & Balleine, B. W. (2007). Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. *The Journal of neuroscience*, *27*(18), 4819-4825.

Ovid, 2009, *Metamorphoses*, A.D. Melville (Trans.). Oxford: Oxford University Press.

Parkinson, J. A., Olmstead, M. C., Burns, L. H., Robbins, T. W., & Everitt, B. J. (1999). Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity byd-amphetamine. *The Journal of neuroscience*, *19*(6), 2401-2411.

Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehnert, B., Lachenal, G., ... & Everitt, B. J. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural brain research*, *137*(1), 149-163.

Parkinson, J. A., Willoughby, P. J., Robbins, T. W., & Everitt, B. J. (2000). Disconnection of the anterior cingulate cortex and nucleus accumbens core impairs Pavlovian approach

behavior: Further evidence for limbic cortical–ventral striatopallidal systems. *Behavioral neuroscience*, *114*(1), 42.

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-870.

Pessoa, L. (2008). 'On the relationship between emotion and cognition.' *Nature Reviews Neuroscience* 9, 148-158.

Pironet, F. and Tappolet, C. (2003). Faiblesse de la raison ou faiblesse de volonté: peut-on choisir?. *Dialogue* 42(04), 627-644.

Plato. (1997). *Protagoras*, in *Complete Works*, J.M. Cooper, D.S. Hutchison (eds.). Indianapolis: Hackett, 746-791.

Prinz, J., & Nichols, S. (2011). Moral Emotions. *The Moral Psychology Handbook*, Ed. John M. Doris. Oxford: Oxford University Press.

Rangel, A., Camerer, C., Montague, P.R. (2008). 'A framework for studying the neurobiology of value-based decision making,' in *Nature Reviews Neuroscience*, available at doi:10.1038/nrn2357.

Ravven, H. M. (2003). Spinoza's Anticipation of Contemporary Affective Neuroscience, in *Consciousness and Emotion* 4 (2), 257-290.

Redish, D.A. (2013). *The Mind Within the Brain.* Oxford: Oxford University Press.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behavioral and Brain Sciences*, *31*(04), 415-437.

Reeve, C. (1992). Introduction. In G. Grube, Plato. *Republic* (revised edition). Indianapolis, Hackett.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Rorty, A. (1980). Where Does the Akratic Break Take Place?. *Australasian Journal of Philosophy* 58: 333-347.

Rorty, A. O. (1980). Akrasia and conflict. *Inquiry: An Interdisciplinary Journal of Philosophy* 23(2), 1980, 193-212.

Rorty, A. O. (1986). Self-deception, akrasia and irrationality. *The multiple self*, 115-31.

Rorty, A. O. (1980). Akrasia and Pleasure: Nicomachean Ethics Book 7. *Essays on Aristotle's Ethics*, 267-84.

Russell, S., Norvig, P. (1995). Artificial Intelligence: A modern approach. *Egnlewood Cliffs*, *Prentice-Hall*.

Ryle, G. (1949/2009). *The concept of mind*. New York: Routledge.

Saarinen, R. (1994). *Weakness of the will in medieval thought: from Augustine to Buridan*. Leiden: Brill.

Saarinen, R. (2011). *Weakness of will in Renaissance and reformation thought*. Oxford University Press.

Sandberg, P. R., Faulks, I. J., Bellingham, W. P., & Mark, R. F. (1981). Relationship between tonic immobility and operant conditioning in chickens Gallus gallus. *Bird Behavior*, *3*(1-2), 51-56.

Santas, G. (1966). 'Plato's Protagoras and Explanations of Weakness of Will.' *Philosophical Review* 75: 3.

Schiffer, S. (1976). A paradox of desire. *American Philosophical Quarterly*,*13*(3), 195-203.

Schmidt, N. B., Richey, J. A., Zvolensky, M. J., & Maner, J. K. (2008). Exploring human freeze responses to a threat stressor. *Journal of behavior therapy and experimental psychiatry*, 39(3), 292-304.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. Neuropharmacology, 37(4), 421-429.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. Cerebral Cortex, 10(3), 272-283.

Schwitzgebel, E. (2010). 'Belief,' *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), available at http://www.science.uva.nl/~seop/entries/belief/

Seth, A.K, Prescott, T.J., Bryson, J.J. (Eds.). (2011). *Modeling Natural Action Selection*. Cambridge: Cambridge University Press.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., & Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664-667.

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience*, 27(18), 4826-4831.

Seymour, B., & Dolan, R. (2008). Emotion, decision making, and the amygdala. *Neuron*, 58(5), 662-671.

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *The Journal of Neuroscience*, 32(17), 5833-5842.

Sheffield, F.D. (1965). Relation between classical and instrumental conditioning. In: W.F. Prokasy (ed.), *Classical Conditioning.* New York, NY: Appleton Century Crofts, 302-322.

Shields, C. (2007). Unified Agency and Akrasia in Plato's Republic. In Bobonich and Destree 2007, 61-86.

Sinnott-Armstrong, W., Mallon, R., McCoy, T, & Hull, J. (2008). Intention, Temporal Order, and Moral Judgments. *Mind and Language* 23: 90-106.

Smith, M., (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion, in Stroud and Tappolet 2003, 17-38.

Spinoza, B. (2002). *Complete Works,* Samuel Shirley (Trans.). Indianapolis: Hackett.

Steiner, A. P., & Redish, A. D. (2014). Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. Nature neuroscience, 17(7), 995-1002.

Stoutland, F. (1968). Basic actions and causality. *The Journal of Philosophy*, 467-475.

Stoutland, F. (1970). The Logical Connection Argument. *American Philosophical Quarterly* 4, 117-129.

Stroud, S. (2014) Weakness of Will. *The Stanford Encyclopedia of Philosophy*.

Stroud, S., and Tappolet, C. (Eds.). (2003). *Weakness of Will and Practical Irrationality*, Oxford: Clarendon Press.

Sullivan, J. P. (1961). The Hedonism in Plato's" Protagoras". *Phronesis*, 10-28.

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human Pavlovian–instrumental transfer. *The Journal of Neuroscience*, 28(2), 360-368.

Tappolet, C. (2003). Emotions and the Intelligibility of Akratic Action, in Stroud and Tappolet 2003, 97-120.

Tenenbaum, S. (1999) The Judgment of a Weak Will. *Philosophy and Phenomenological Research* (59), 875-911.

Thagard, P. (2010). *The Brain and the Meaning of Life*. Princeton, NJ: Princeton University Press.

Thagard, P., & Kroon, F. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge: MIT Press.

Thaler, R.H., Sunstein, C.R. (2008). *Nudge: Improving Decisions About Health, Wealth and Happiness*. New York: Penguin.

Thero, D. (2006). Understanding Moral Weakness, Amsterdam, New York: Rodopi.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, *96*(6), 3257-3262.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. Ethology and sociobiology, 11(4), 375-424.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. Nature, 398(6729), 704-708.

Tremblay, L., & Schultz, W. (2000). Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*(4), 1864-1876.

Tremblay, L., & Schultz, W. (2000). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of neurophysiology*, *83*(4), 1877-1885.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29(11), 2225-2232.

Tyrrel, T. (1993). *Computational Mechanisms for Action Selection*. (Doctoral Dissertation). University of Edinburgh, Edinburgh.

Valentin, V.V., Dickinson, A., O'Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Human Neuroscience* 27: 4019-4026.

Vlastos, G. (1956). Introduction. *Plato's Protagoras*. New York: Library of Liberal Arts, xxxviii-xlv.

Vlastos, G. (1969). Socrates on Akrasia. *Phoenix* 23: 71.

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological science*, *11*(3), 249-254.

Walker, A. (1989). The Problem of Weakness of Will. *Noûs* 23: 653-676.

Watson, G. (1977). Skepticism About Weakness of Will. *Philosophical Review* 86: 316-339.

Wegner, D. M. (2002). *The illusion of conscious will*. MIT press.

Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends in cognitive sciences*, *7*(2), 65-69.

Weinstein, I. B., & Case, K. (2008). The history of Cancer Research: introducing an AACR Centennial series. *Cancer research*, *68*(17), 6861-6862.

Weiss, R. (2007). Thirst as Desire for Good. In Bobonich and Destree 2007, 87-100.

Wiggins, D. (1979). Weakness of Will, Commensurability, and the Objects of Deliberation and Desire. *Proceedings of the Aristotelian Society* 79: 251-277.

Wilkerson, T.E. (1997). *Irrational Action: A Philosophical Analysis*. Aldershot: Ashgate.

Williams, D.R., & Williams, H. (1969). Auto-maintenance in pigeon – sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior* 12 (4), 511-520.

Wyma, K. D. (2004). *Crucible of reason: intentional action, practical rationality, and weakness of will*. Rowman & Littlefield.

Yam, K. L. (2009). *The Wiley Encyclopedia of Packing Technology*, 3rd Edition. Hoboken, NJ: Wiley.