An Empirical Solution to the Puzzle of Weakness of Will

Julia Haas

Section 1 – Introduction

Weakness of will occurs when an agent holds that it would be better not to do some action, but does it anyway.[1] For example, Gene is weak willed when he thinks it would be better not to have cake, but nonetheless decides to eat some.[2] The experience is often accompanied by feelings both of having acted freely and of having given in to temptation (Sripada 2010).[3]

So described, the *phenomenon* of weakness of will is difficult to explain, but not in itself paradoxical. The *puzzle* of weakness of well is the product of theoretical assumptions implicit (and sometimes explicit) in folk and philosophical psychology about the nature of evaluative judgments and their role in decision making. The puzzle can be captured as follows:

FOLK PSYCHOLOGICAL THEORY [4]	If, at time *t*, an agent judges that it is better to do A than B, and he believes he is free to do A, then, provided he tries to do either at that time, he will try to do A and not B.[5]

[1] Weakness of will and *akrasia* are used interchangeably throughout (although see Holton 1999, 2009; Mele 2010).

[2] Following Rorty (1980), the so-called akratic break can occur at various points, including both deciding to eat the cake and eating it.

[3] The opposing feelings of giving in and of acting freely, or 'uncompelledness,' are sometimes jointly described as a feeling of *inner conflict*. E.g. Hare describes weakness of will as a psychological circumstance best expressed by the "curious metaphor of divided personality which, ever since this subject is first discussed, has seemed so natural" (1963, 81).

[4] Folk psychological theories are here philosophical theories that aim to explain choice and action using (purportedly) commonsense assumptions about judgments, beliefs, desires, etc. (Lewis 1972, Stich and Nichols 2003). FOLK PSYCHOLOGICAL THEORY is broadly representative of those theories as presented in the context of explaining weakness of will (see Stroud and Tappolet 2003 for a review). Weakness of will is a central test case for such theories, since they have difficulties explaining how the phenomenon could be possible. My goal in this paper is to provide a data-driven rather than a folk psychological theory of weakness of will.

[5] Per the recommendations of two anonymous referees, FOLK PSYCHOLOGICAL THEORY and WEAKNESS OF WILL accommodate the temporal features of choice, the principle that an agent need only *believe* that she is free to act in a certain way, and unsuccessful attempts at action. Although both FOLK PSYCHOLOGICAL THEORY and WEAKNESS OF WILL are still vulnerable to additional counterexamples, they can, in principle, be revised to accommodate them. My goal is not to establish a theory of weakness of will that precludes all possible counterexamples, however. My goal is to provide a plausible, i.e., empirically-informed theory of weakness of will that nonetheless does not result in a paradox. See also Ft. 6 below.

WEAKNESS OF WILL	In cases like Gene's above, an agent judges that it is better to do A than B, believes that he is free to do A, but tries to do B.[6]

The two statements are inconsistent: FOLK PSYCHOLOGICAL THEORY precludes the possibility of weakness of will (as characterized in WEAKNESS OF WILL), but WEAKNESS OF WILL asserts that it occurs.

This paper presents an empirical solution to the puzzle of weakness of will. Specifically, it presents a theory of action, grounded in contemporary cognitive neuroscientific accounts of decision making, that explains the phenomenon of weakness of will without resulting in the puzzle. Thus, it revises FOLK PSYCHOLOGICAL THEORY and thereby accommodates WEAKNESS OF WILL.

I'll place my cards on the table. I'm arguing for a rather modest conclusion to a classic philosophical debate: the phenomenon of weakness of will is possible and explainable by ordinary features of our decision-making systems. The puzzle, as expressed by the inconsistent statements above, is incoherent. The philosophical debate surrounding weakness of will is motivated in part by equating the *phenomenon* of weakness of will with a particular folk psychological *interpretation* of that phenomenon. Employing a different interpretation enables a novel understanding of weakness of will.

Section 2 – Standard discussions of weakness of will

There have been two types of responses to the puzzle: (1) To re-describe the statements, and thereby show it to be only apparently contradictory; or (2) to supplement either FOLK PSYCHOLOGICAL THEORY or WEAKNESS OF WILL with ancillary mental states, and thereby remove the inconsistency.[7] Early responses to the puzzle were of the first type. Donald Davidson (1970) argued that reading FOLK PSYCHOLOGICAL THEORY as involving

[6] WEAKNESS OF WILL does not define weakness of will. It identifies a class of actions characterized by a sufficient condition. I have tried to choose as neutral a sufficient condition as possible, though of course my choice of condition is still controversial. For example, WEAKNESS OF WILL is often formulated using the notion of "intending to do," rather than "trying to do." I prefer the latter formulation, since the notion of intentions as distinct mental states is facing pressure from the cognitive neurosciences (for an excellent discussion of using intentions in neuroscientific explanations, see Uithol, Burnston, and Haselager 2014). However, my theory could be modified to accommodate intentions as well as alternative sufficient conditions. At present, my theory only addresses those cases of weakness of will captured by WEAKNESS OF WILL. I thank two anonymous referees for helping me clarify this point.

[7] A third response has been to deny that weakness of will ought to constrain our analysis of practical reasoning (e.g. Hare 1952, 1963).

'unconditional' evaluative judgments and WEAKNESS OF WILL as involving 'conditional' evaluative judgments rendered the statements consistent.[8]

Unconditional judgments are general assessments about what is better or best, independently of any further considerations, taking the form:

> UJ: *a* is better than *b*.

Conditional judgments are made relative to specific factors and considerations, taking the form:

> CJ[1]: in light of consideration(s) *c, a* is better than *b* (Davidson 1970, 37–42; see also Stroud and Tappolet 2003, 3-4).

Unconditional judgments purportedly result in intentional actions, whereas conditional judgments only arrive at theoretical conclusions about what it would be best to do. As a result, an agent can arrive at the conditional evaluative judgment that, in light of certain considerations, '*a* is better than *b*,' but still do *b*. In our example, this means that if Gene had made an unconditional judgment that it would be better not to eat the cake, then he could not do so. Since he made a conditional judgment about what would be better, it remains possible for him to do so. Thus, Davidson suggests, Gene is not enacting "a simple logical blunder" by being weak-willed, although his actions may be considered irrational (1970, 40).

Philosophers have criticized Davidson for failing to solve what some see as the genuine puzzle of weakness of will, namely, weakness of will involving unconditional judgments (see Stroud and Tappolet 2003 for a review). In contrast to garden-variety weakness of will, these cases can be called instances of *unconditional weakness of will*.[9]

It is increasingly recognized that the problem of unconditional weakness of will cannot be resolved by re-describing the statements (Bratman 1979, Davidson 1982, Pears 1984, Mele

---

[8] Conditional evaluative judgments are also called *prima facie* or *all things considered* judgments. Unconditional evaluative judgments are also called *all-out* judgments about what it would be best to do (Davidson 1970).

[9] This type of weakness of will is also known as last-ditch akrasia (Pears 1984), strict akratic action (Mele 1987), and clear-eyed weakness of will (Bobonich and Destree 2007). In a representative passage, Robert Dunn argues,

> Davidson too is revealed as unsympathetic to the possibility of [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 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).

1987, Dunn 1987, Buss 1997; though see Stroud 2003). Instead, many have turned to the second approach, namely, of supplementing FOLK PSYCHOLOGICAL THEORY or WEAKNESS OF WILL with ancillary mental states. Supplements have included the introduction of evaluative commitments (together with a principle of rationality) (Bratman 1979), intentions (Audi 1979, Holton 1999), moods (Stocker 1979), and emotions (Tappolet 2003); distinguishing between evaluating and wanting (Dunn 1987, Mele 1987); and the turn to empirically derived concepts such as ego depletion (Levy 2011). These accounts revise the theoretical architecture driving the problems of weakness of will in general, and unconditional weakness of will in particular (Rorty 1980, Davidson 1982).

Few philosophers have sought to resolve the puzzle of weakness of will by explicitly committing to a multi-system model of the mind, on which two or more systems operate in parallel to produce action (Plato 1997 (*Protagoras),* Aristotle 1985 (*Nicomachean Ethics* 7.1-10), Augustine 1960 (*Confessions* 8.5), Aquinas 1952 (*Summa Theologica* I-II q. 77, art. 1 and 2, II-II q. 156, art. 1), Spinoza 2002 (*Ethics* Part IV), Leibniz 1965 (*New Essays on Human Understanding* II, ch. 21, 35), Davidson 1970, Bratman 1979, Audi 1979, Stocker 1979, Rorty 1980, Dunn 1987, Mele 1987, 1992, Buss 1997, Holton 1999; although see Davidson 1982, Pears 1984, Plato 1997, Sripada 2010, and Levy 2011 as key exceptions, below). Not committing to such a multi-system model makes weakness of will difficult to explain, however, since some ancillary state must be said to disrupt the standard progression toward action. For example, if an agent arrives at an unconditional judgment that it would be best to do *a*, but she does ~*a,* then some ancillary attitude or event, such as a wayward desire or wanting, must be said to disrupt the normally reliable transition from unconditional judgment to corresponding action. The introduction of such ancillary states leaves single-system accounts vulnerable to the charge that they provide merely ad hoc solutions to the puzzle of weakness of will. As Alfred Mele notes,

> When irrational behaviors cannot be stretched or cut to fit a favored model, their very existence may be denied. And when they are made to fit, they often look more rational than one should have thought. Sometimes, of course, the *models* bend; but not far enough, or not in the right places" (1987, vii, emphasis original).

Conversely, such attitudes may go too far, since they are said to not only cause, but also to provide what amounts to a preponderance of reasons for what would otherwise be considered instances of weakness of will. But if an agent judges that she has a preponderance of reasons for acting in a certain way, her action can no longer be described as weak-willed (Davidson 1982).

In response to such difficulties, some have proposed to resolve the paradox by positing a partitioned or multi-system model mind (Davidson 1982, Pears 1984; see also Sripada 2010). On *multi-system* models, as I will call them, the mind can issue simultaneous but inconsistent judgments and actions. For example, Davidson (1982) proposes that two or

more autonomous mental structures issue unconditional judgments and actions independently. On this view, some unconditional judgments may cause actions *without* providing reasons for them, and thereby secure the possibility of unconditional weakness of will. Davidson offers the analogy:

> Wishing to have you enter into my garden, I grow a beautiful flower there. You crave a look at my flower and enter my garden. My desire caused your craving and action, but my desire was not a reason for your craving, nor a reason on which you acted. (Perhaps you did not even know about my wish) (1982, 181).

Just as Davidson's desire causes but is not the reason for your entering the garden, so one mental event can cause another without providing a reason for it.

Traditional multi-system accounts use the occurrence of weakness of will as a reason for positing multiple systems in the mind. But this approach makes it difficult for such accounts to provide *independent* reasons for thinking that the mind is actually partitioned in that way. Consequently, multi-system models of the mind have been unusually poorly received. They are criticized as being vague (Mele 1987), unnecessarily complex (Heil 1989), and empirically implausible (Peijnenburg 2000).

Since critiques of multi-system views have highlighted issues of plausibility and support, the debate would benefit from a principled, domain-general multi-system model of the mind. Such a model could explain both successful practical reasoning and less ideal cases such as instances of weakness of will. It would also provide empirical evidence in support of its central explanatory features, such as the existence of multiple mental systems, and thereby better avoid the charge that these features are merely theoretically convenient.

The remainder of this paper offers such a multi-system solution to the puzzle of weakness of will. The next section lays out the principles of a multi-system model of the mind. Section 4 shows how multiple systems explain the phenomenon of weakness of will.

Section 3 – A Multi-System Model of the Mind

Research in reinforcement learning, neuroeconomics, and cognitive neuroscience has made substantial contributions to our understanding of deliberation and choice (Sutton and Barto 1998, Glimcher 2010). The present account draws on these disciplines, henceforth "the decision sciences," to describe what I call the Multi-System Model of the mind (MSM).

Describing an agent as capable of deliberation and choice involves making basic assumptions about her mental capacities. Standard philosophical approaches describe these capacities in terms of practical reasoning, asking how an agent deliberates about what to do, and how she can act accordingly. Such approaches answer these questions using folk psychology which, as in FOLK PSYCHOLOGICAL THEORY, aim to explain choice and action using commonsense assumptions about judgments, beliefs, and so on. (Lewis 1972, Stich and Nichols 2003).

The decision sciences recast the problem in terms of *learning* (Barto 1995). An early theory proposes that learning only occurs when an event violates an agent's expectations (Rescorla and Wagner 1972; see also Schultz et al.1997). For example, once animals learn to associate a sound and an event, they can no longer learn to use a second stimulus, such as a light, to predict the same event.[10] A later theory, introduced by Richard Sutton and Andrew Barto (1998), adapted the so-called Rescorla-Wagner learning rule to analyze how agents learn to make more efficient decisions over time.

According to Sutton and Barto (1998), every agent aims to maximize value (see also O'Doherty 2014, Kable and Glimcher 2009). Value is a function of an agent's rewards and punishments over time, and is calculated subjectively by the agent (Padoa-Schioppa and Assad 2006). An agent aims to secure reward and avoid punishment; valuation thus underwrites motivation. But, as we shall see, the learning mechanisms ensuring the overall maximization of value mean that valuation does not directly underwrite action.

Sutton and Barto (1998) propose that an agent uses learned experience to maximize value. To maximize value, an agent must accurately and effectively predict future value.[11] To predict future value, an agent uses one or more computational strategies to assess different courses of action. The major task and achievement of research in reinforcement learning over the past several decades has been to elucidate specific computational strategies that make these kinds of assessments possible. Following consensus in the decision sciences, MSM posits multiple computational strategies that underwrite and define semi-independent decisions systems (for two excellent introductions, see Montague 2007 and Redish 2013; for more specialized reviews, see Dayan and Niv 2008, Rangel *et al.* 2008, Dayan 2011, Glimcher and Fehr 2013).[12]

[10] I.e., Kamin Blocking (Kamin 1968).

[11] *Securing* the relevant value alternatives represents an associated challenge. It is of no use to predict that the juiciest and most nutritious leaves are on the highest branches of the tree if one cannot also reach those highest branches.

[12] While most evidence suggests that decisions are controlled by at least three distinct systems, the full range of decision-making behaviors may be underwritten by still more. For example, there may be more than one type of Pavlovian controller (see Daw and O'Doherty 2013 for a recent discussion of this issue). Thus, while the *Multi*-System Model refers to just three systems in practice, it allows for still more in theory.

The *hardwired* system represents the most basic of the multiple systems. Behaviors issued by this system are characterized by automatic approach and withdrawal responses to appetitive and aversive stimuli, respectively (Macintosh 1983).[13] When Gene sees cake, he tends to approach it. When he sees a snake, he tends to withdraw from it. Hardwired responses are appropriate in natural environments, since it is broadly beneficial to approach rewards and avoid punishments. However, hardwired responses also lack flexibility, which can result in detrimental decision outcomes, as we shall see below (Huys *et al.* 2012).

Because of their behavioral rigidity, researchers identify hardwired responses by persistence even in those cases where persistence is detrimental. For example, David and Harriet Williams demonstrated that pigeons continue to peck at a key associated with food, even when doing so resulted in losing the reward (1969, see also Brown and Jenkins 1968). In doing so, Williams and Williams provided evidence for the existence of an automatic, hardwired decision system. Similar results have been found with other animals (Sheffield 1965, Herschberger 1986, Bouton 2006). Hardwired behaviors have also been identified in human participants (Redish 2013).[14]

A second, *deliberative* system explicitly represents possible choices and determines the sequence of actions that maximizes value.[15] This procedure is typically represented by a decision tree: each node in the tree representing a possible choice. Total value is the sum of the rewards minus the sum of the punishments along a given branch. The deliberative system searches through the decision tree to find the branch with the highest total value. Hence its alternative name: tree search. For example, Gary Kasparov may represent three or four upcoming moves in a game of chess, with each possible move further branching into a wide range of subsequent moves. To win, Kasparov tries to represent and choose the best possible sequence of moves overall.[16]

However, as Peter 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 on the order of 30" (2011). Whereas

[13] In reinforcement learning, this system is formally called the Pavlovian system (e.g., see Dayan *et al.* 2006, Talmi *et al.* 2008, Dayan and Berridge 2014). However, the term 'Pavlovian' frequently leads to confusion among researchers in other fields (for an interesting discussion of how psychologists and machine learning scientists characterize the Pavlovian system differently, see Rescorla's "Pavlovian conditioning: It's not what you think it is," 1988). In most fields, as well as in everyday usage, the term 'Pavlovian' is usually associated with Pavlov's original experiments with dogs, where Pavlov trained his dogs by repeatedly ringing a bell and then consistently feeding them afterwards. By contrast, in reinforcement learning, *it is the relationship between the unconditioned stimulus (i.e., the food) and the unconditioned response (i.e., the salivating) that is of interest.*

[14] Basic emotional responses, such as fear and anger, are considered hardwired responses (Redish 2013; see also Barrett 2006). The hardwired system is also increasingly thought to influence impulsivity (Ainslie 2001), anxiety, and depression (Dayan and Huys 2008, Huys et al. 2011, Huys et al. 2012).

[15] The deliberative system is formally called the goal-directed or model-based system (e.g. see Dayan 2011).

[16] Example from Dayan (2011).

hardwired responses are automatic and inflexible, the deliberative system responds flexibly to new situations, but struggles if there are many options to consider. As a result, agents employing the deliberative system 'prune' their decision trees, focusing on promising branches of the decision tree and ignoring less obviously advantageous ones (Huys *et al.* 2012). Since some branches may initially appear to be disadvantageous only to become advantageous later, pruning can prevent an agent from finding the best sequence of moves overall.

Unlike the deliberative system, a third, *habitual* system does not explicitly represent future alternatives, but caches positive and negative experiences, and assigns values to actions based on their previous outcomes.[17] Good state-action pairs have produced rewarding outcomes in the past, and so should be repeated. Bad state-action pairs have produced punishments in the past, and so should be avoided.[18] To cache experiences, the habitual system employs a feedback signal, which revises the system's estimates about the environment.[19]

Because of their distinct methods of assessing value, the different systems provide more and less accurate predictions of value in different contexts. Due to its fast, automatic nature, the hardwired system has a high probability of arriving at an accurate prediction in a simple, single-answer problem, such as approaching reward. Due to it flexible, forward-looking nature, the deliberative system has a high probability of arriving at an accurate prediction in an unfamiliar environment. And in virtue of its efficient caching procedure, the habitual system has a high probability of arriving at an accurate prediction in a complex, familiar context.[20]

---

[17] The habitual system is formally called the habit-based or model-free system (e.g. see Dayan 2011). Experimental psychologists have dissociated deliberative- and habit-based activities in animals. Anthony Dickinson and Bernard Balleine trained rats to press a lever in exchange for a reward, and then devalued the reward by pairing it with a noxious substance. They then examined whether the animals continue to press the lever to receive further rewards. Notably, the duration of the rats' initial training determined whether they were willing to press the lever or not. If they were trained for a moderate period of time, the rats no longer pressed the lever. If they were trained for a longer period, the rats continued to press the lever. These responses have been interpreted as reflecting the deliberative system in the first case and the habitual system in the latter case (Dickinson 1985, Dickinson *et al.* 2002). Modified versions of this methodology have been used to isolate deliberative activity in human participants (Hampton *et al.* 2006, Valentin *et al.* 2007, Tricomi *et al.* 2009).

[18] I owe this formulation to Crockett 2013.

[19] The feedback signal works much like exclamations of 'Hotter!' and 'Colder' in the children's game Hot-or-Cold. The Seeker moves around the room with the general goal of finding a hidden object. The Hider helps the Seeker by telling her whether she is getting closer or farther away. The Hider's suggestions operate like an error signal by helping the Seeker refine her predictions, albeit without giving her detailed instructions about where to go (analogy from Montague 2006).

[20] Taking the usual subway route home, driving a car, playing a memorized musical composition, and completing a practiced play in football are all examples of complex, familiar decision tasks (see Redish 2013, Chapter 10; see also Cisek and Kalaska 2010).

Having multiple systems competing to evaluate a single task might seem at best inefficient and at worst disadvantageous. However, simulations suggest that optimal agents employ several complementary controllers (Dayan 2011, Daw *et al.*, 2005, Lee *et al.* 2014). As Nathaniel Daw and colleagues note, "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 [system] of each class in circumstances in which predictions tend to be most accurate" (2005, 1704).

Two competing models explain arbitration between the systems.[21] On the 'hierarchical' model of arbitration, interactions between the systems are governed by the principle:

> HA The deliberative system has veto power over the hardwired and habitual systems.

HA intimates that the deliberative system exercises something akin to top-down cognitive control, including control of the hardwired and habitual systems (Cohen *et al.* 1996, Botvinick *et al.* 2001). For example, HA explains how an agent can exercise control and avoid pursuing a tempting but ultimately costly alternative.

On the alternative model, interactions between the systems are governed by the following 'accuracy-based principle of arbitration':

> PA Following partial evaluation, that system with the highest *accuracy profile*, i.e., that system most likely to provide an accurate prediction of expected value, relative to the decision problem at hand, directs the corresponding assessment of value (Deneve and Pouget 2002, Daw *et al.* 2005, Lee *et al.* 2014).[22]

On this model, each of the multiple systems partially evaluates the action alternatives. Simultaneously, each system generates an estimate of how accurate its prediction is relative to decision problem at hand. These estimates, or accuracy profiles, are then compared, and the system with the highest accuracy profile is selected to direct the corresponding valuation task. PA thus directs valuation based on a system's accuracy profile, and not on its prediction of value. For example, the hardwired system typically coordinates choice in familiar, complex settings, because it typically has a higher accuracy profile in those decision problems, even if it predicts a lower overall value than do either its deliberative or

[21] Thanks to an anonymous referee for raising the issue of hierarchical models, and for encouraging me to clarify my discussion of PA, below.

[22] Partial evaluation further optimizes choice, "trading off the likely costs (for example, time or calories) of additional search against its expected benefits (more accurate valuations allowing better reward harvesting)" (Daw *et al.* 2005, 1708).

hardwired counterparts. Crucially, PA ensures an agent can assess an action A as being preferable to action B, but still do B.

MSM follows the PA solution initially proposed by Daw *et al.* (2005), for three reasons. First, while hierarchical models of control are popular in the cognitive science literature (Badre 2008, Botvinick *et al.* 2009, Uithol *et al.* 2012, Clark 2013, Hohwy 2013), the only precisely characterized version of such a view in the narrower domain of reinforcement learning is presented in Rangel (2013, 5-6).[23] It proposes:

> HA1 The deliberative system overrides the hardwired and habitual systems *when it is beneficial to do so*, e.g., when it computes the correct overall value of a given choice, detects a conflict between this value and the values computed by the other systems, and inhibits their competing responses.

But this view faces a challenge. It is not clear how, on HA1, the deliberative system could establish when it would be beneficial to override its counterparts, e.g., how it could establish that *it* has calculated the 'correct' overall value. The deliberative system could not, for example, assume that it is correct whenever it has calculated the highest value overall, since one of its central computational advantages lies in its ability to devalue short-term rewards (Hare *et al.* 2009). Hence, HA1 does not explain how the deliberative system would 'know' to exercise top-down cognitive control over the hardwired and habitual systems.

Nonetheless, this criticism of HA1 does not demonstrate that the more general principle, HA, is problematic. A second, empirical consideration weighs in favor of using PA rather than HA for the current project: PA represents the canonical approach in the reinforcement learning literature, and it is strongly supported by both behavioral (see Dickinson and Balleine 2002 for a review) and neuroanatomical evidence in both non-human and human animals (Balleine and Dickinson 1998, Balleine and Dickinson 2000, Killcross and Coutureau 2003, Izquierdo *et al.* 2004, Kiani and Shadlen 2009, Lee et al. 2014). Thus, while HA remains an empirical possibility, its prospects are not as good as those of PA.

The final reason for using PA is a pragmatic one. For a model of weakness of will, it is of crucial importance to determine how and when the different systems exercise control over choice and action. As noted above, PA provides a detailed and predictive account of such interactions. But HA, as characterized in general terms, does not specify the conditions

[23] See my discussion of Levy (2012) in Section 5, below, for an example of a hierarchy-based explanation of weakness of will that is applied outside the context of reinforcement learning.

under which the deliberative system has veto power over the other systems or, more generally, when it would exercise it. MSM thus adopts PA as its principle of arbitration.[24]

Taking stock: the MSM equivalent of FOLK PSYCHOLOGICAL THEORY from Section 1, expressed in terms of maximizing value, states:

FOLK PSYCHOLOGICAL THEORY*	If, at time *t*, an agent's decision system D values some option A more highly than some option B, and he believes he is free to do A, then, provided he tries to do either at that time, he will try to do A and not B.

But PA entails that FOLK PSYCHOLOGICAL THEORY* is false. Instead, MSM proposes:

MSM THEORY	If, at time *t*, an agent's decision system D values some option A more highly than some option B, he believes he is free to do A, *and* PA allocates D for action-selection, then, if he tries to do either at that time, he will try to do A and not B.

Thus, on MSM THEORY, an agent may judge that it is best to do A rather than B (in that one of his decision systems values A more than B) and yet do B, if that system is not the one allocated for action selection. In the next section I show how this insight explains weakness of will.

Section 4 – Explaining Weakness of Will

In the history of philosophy, interest in weakness of will focused on how to formulate FOLK PSYCHOLOGICAL THEORY, given the widespread acceptance of WEAKNESS OF WILL, and not to render the statements consistent. That is, they attempted to explain the phenomenon and not the puzzle. For example, in the *Protagoras,* Plato's Socrates finds the phenomenon so familiar that he uses it to illustrate a broader claim about virtue, but

[24] That said, if HA does turn out to be correct and, further, provides the relevant conditions for the deliberative system enacting control, the model of weakness of will presented here should apply, *mutatis mutandis*, to HA.

disagrees with the commonsense explanation of what causes it (353a).[25] Similarly, in the *Republic,* Plato accepts the everyday phenomenon but now explains it using the tripartite soul (448-444b; see also Aristotle 1985 (*Nicomachean Ethics* 7.1-10), Augustine 1960 (*Confessions* 8.5), Aquinas 1952 (*Summa Theologica* I-II q. 77, art. 1 and 2, II-II q. 156, art. 1), Spinoza 2002 (*Ethics* Part IV), Leibniz 1965 (*New Essays on Human Understanding* II, ch. 21, 35)). MSM provides this kind of explanation. On MSM, interactions between the multiple systems account for both sound decision-making and weakness of will. As I will show, these interactions elicit at least three distinct types of weakness of will, which I call *habitual*, *pruning-based*, and *inhibitory*.

Habitual weakness of will is elicited by interactions between the deliberative and habitual systems. Recall, the deliberative system typically has a higher reliability measure in novel settings, since its capacity for representation allows it to predict the values of various outcomes. By contrast, the habitual system is typically more reliable in complex but familiar circumstances, where representation would be both taxing and redundant. But it is not unusual for an important aspect of a familiar situation to change. Such circumstances elicit the most basic and harmless type of weakness of will.

If an agent opts for the typically more reliable but in fact inaccurate habitual system, she experiences habitual weakness of will. The information provided by the deliberative system enables the agent to know what the best course of action would be under these recently changed circumstances. Yet since the situation is broadly familiar, the habit-based approach has a high past cumulative success rate, or an overall high reliability measure. Thus, PA allocates the habitual system for action selection. Hence, the agent is aware of the most up-to-date and appropriate course of action in advance, but falls back on her less beneficial, habitual counterpart (for an extended discussion, see Daw *et al.* 2005). The agent experiences the signature phenomenology of weakness of will: she recognizes that it would be preferable to do A, but feels herself choosing to do B.[26]

Habitual weakness of will accounts for several paradigm cases of weakness of will, including Davidson's classic example of brushing his teeth. 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 gets out of bed and goes to brush his teeth anyway (1970, 30). Here, MSM makes sense of the otherwise perplexing action: the habitual system dictates that brushing one's teeth is a

[25] The rejection of the multitude's view is sometimes interpreted as Socrates' denying the occurrence of the everyday experience altogether, or the 'Socratic denial of akrasia' (e.g. Walsh 1963). This is mistaken (see Penner 1997, Shields 2007 for two discussions of this issue).

[26] As it is the agent's own habitual system that computes and assigns the relevant values to B, there is no reason to expect that she will feel compelled or otherwise 'un-free' in making her decision.

reliably valuable course of action, even though the circumstances make it the less valuable action overall.[27]

In *pruning-based weakness of will*, the deliberative and hardwired systems interact to issue a suboptimal choice. This second type of interaction occurs when an option represented in the deliberative system's decision tree elicits either a positive or negative hardwired response. For example, a strongly positive alternative, represented at an early node of the decision tree, may cause the entire opposing branch of the tree to be pruned, so that it is no longer considered. Conversely, a strongly negative alternative, represented at an early node of the decision tree, may cause the entire subsequent branch of the tree to be pruned, so that none of the subsequent values are computed or represented (Huys et al. 2012). In principle, such a pruning mechanism combines the virtues of both representation and efficiency to narrow a potentially intractable computational problem. But recall from Section 3 that in some cases, pruning can prevent an agent from considering alternatives that are most beneficial overall. Pruning is detrimental in those cases where a strongly positive or negative alternative masks a more gradual but substantial accumulation of value further down in the tree.[28]

We can recast the original example of Gene considering the cake in terms of pruning-based weakness of will. Gene represents his options in the form of a decision tree, consisting of both 'eat' and 'don't eat' alternatives. The highly appealing nature of the cake, represented early in the tree, engages the hardwired system and causes the branch representing the possibility of *not* eating the cake to be pruned. Gene thus represents the negative consequences of the remaining branch – the negative consequences of eating the cake - but no longer has any other alternative available to him. So, he eats the cake.[29]

Pruning-based weakness of will can similarly account for another classic case of weakness of will, described by J.L. Austin:

---

[27] Since the habitual system continues to re-evaluate its choices as it acquires new experiences, habitual weakness of will promises to be remediable. That is, although Davidson might get up to brush his teeth once or twice, he would experience the action's negative consequences and refrain from repeating it in the future. Interactions between the deliberative and habitual systems thus provide one explanation of the everyday experience of weakness of will.

[28] In a recent experiment, Huys and colleagues showed that when one of the branches of a decision tree is associated with a large loss early in the tree, participants rely on the hardwired system to prune out that entire series of actions (Huys et al. 2012).

[29] One could object at this point that, per WEAKNESS OF WILL, this is not really a case of weakness of will. But Gene may still be aware of his pruned options. What the agent consciously perceives as his options may not match what happens at the level of his decision systems. Thanks to an anonymous referee for pressing this point.

> 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).

Austin begins to represent his options in the form of a decision tree, consisting of the 'eat' and 'don't eat' alternatives. The full tree would represent the ensuing consequences of both alternatives. But the highly appealing nature of the bombe, represented early in the tree, engages the hardwired system and causes the tree to be pruned. Austin thus represents the negative consequences of the only remaining alternative – eating the bombe – but has no other alternatives left to pursue. He eats the bombe. The specific phenomenal features of Austin's experience are also accounted for. Although it is the product of the hardwired system, pruning-based weakness of will needn't be rushed or impulsive. Rather, the pruning of the decision tree simply eliminates certain choices, and thereby leaves the less optimal alternative to be pursued "with calm and even with finesse."[30]

Pruning-based weakness of will involving *negative* options proceeds analogously. The Milgram obedience experiments studied participants' willingness to obey an authority figure, known as EXPERIMENTER. Participants were asked to administer shocks to an alleged fellow participant, known as LEARNER.[31] In the baseline condition of the study, many participants obeyed EXPERIMENTER until the end, punishing the supposed victims with the strongest shock possible (Milgram 1963).

Pruning-based weakness of will can help explain these behaviors. The participants faced a strongly negative alternative at the very top of their decision trees: an immediate confrontation with EXPERIMENTER. Moreover, EXPERIMENTER commanded the participants to administer the increasingly severe shocks, repeating a series of "prods," such as, 'It is absolutely essential that you continue,' etc. (Milgram 1974, 21-22). The prospect of this intensely unpleasant interaction may have pruned the participants' decision tree, so that

[30] As noted by an anonymous referee, it may be objected that Gene's case and Austin's example amount to instances of inhibitory rather than pruning-based weakness of will. But inhibitory weakness of will is characterized by either physical immobility or nearly compulsive activity, and neither Gene nor Austin experience either of these (e.g., Austin does not 'raven'). Their cases are thus best understood as instances of pruning-based weakness of will.

[31] The participants' behaviors in the Milgram studies suggest that they correspond to instances weakness of will. Even those participants who continued to administer the shocks did so under extreme stress. For example, many participants perspired heavily, laughed at inappropriate times, and even experienced seizures (Milgram 1974). In their analysis, Merritt *et al.* (2011) interpret the participants' acute symptoms of distress as indicating that the participants did not endorse the violent punishment of the victim, but continued to press the button anyway.

the 'stop shocking' course of action no longer seemed available. In the false binary of the experimental set-up, this left only the alternative of shocking the alleged fellow participant.

This interpretation is supported by findings from several of the experiment's 18 variations. For example, in Experiment 14, EXPERIMENTER is placed in the role of LEARNER, and a less authoritative figure instructs the naïve participant to shock LEARNER. Milgram writes, "[the less authoritative figure's] instructions to shock [EXPERIMENTER] were totally disregarded... At the first protest of [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).[32] In this variation, the less authoritative figure represents a lesser threat to the participant. The participant's decision-making alternatives remain 'open' (i.e., unpruned), and disagreeing with EXPERIMENTER remains on the table.

Inhibitory weakness of will occurs when an agent's deliberative system identifies the best course of action, but the hardwired system physically inhibits it (Huys et al. 2012). The hardwired system can overrule the deliberative system when negative associations discourage an agent from pursuing a goal, or positive associations encourage an agent to pursue what would otherwise be considered an aversive outcome. For example, if an athlete hears a sad song during her run, she may begin to feel slow and tired before reaching her destination.[33] Similarly, pleasant company or having a glass of alcohol may encourage a smoker to yield to the temptation of a cigarette.[34] In both cases, the agent recognizes the options available to her, but is moved to pursue the inferior action. The agent thus experiences the signature phenomenology of weakness of will: recognizing that it would be better to do A than to do B, but nonetheless doing B.

Inhibitory weakness of will accounts for Mele's Rocky case. Mele describes the example:

> Rocky, who has promised his mother that he would never play tackle football, has just been invited by some older boys to play in tomorrow's pick-up game. He believes that his promise evaluatively defeats his reasons for playing and consequently judges that it would be best not to play; but he decides to play anyway. However, when the time comes, he suffers a failure of nerve. He does not show up for the game—not because he judges it best not to play, but rather because he is

[32] Along slightly different lines, in Experiment 15, the baseline condition remained the same, but there were two EXPERIMENTERS in the room with the participant instead of one. When LEARNER protested at the shock, the two EXPERIMENTERS verbally disagreed 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).

[33] The hardwired system can also support or enhance the deliberative system. For example, if an athlete deliberately runs up a challenging hill, hearing her favorite song triggers her hardwired system to subconsciously picks up the pace.

[34] Thanks to [blinded] for this example.

> afraid. He would not have played even if he had decisively judged it best to do so (1987, 7).

Although it is hard to understand Rocky's situation using FOLK PSYCHOLOGICAL THEORY, the inhibition mechanism helps illuminate his actions. The inhibitory mechanism overrides Rocky's deliberation-based assessment, independently of his promise to his mother. Watching the others play football triggers Rocky's hard-wired fear response, preventing him from joining in. Inhibitory weakness of will thus accounts for even those rare cases of weakness of will that *align* with an agent's assessment of what she should do (see also Arpaly 2000).

Thus, MSM accounts for the phenomenon of weakness of will, without resulting in a puzzle. Moreover, as developed in the next section, MSM has the advantage of identifying multiple dissociable kinds of weakness of will.

Section 5 – Solution and Implications

In Section 1, I stated that a successful solution explains the phenomenon of weakness of will without arriving at the puzzle represented by the inconsistent statements. MSM provides such a solution. As outlined in the previous section, MSM explains the phenomenon by describing the mechanisms underlying the signature phenomenology and contradictory actions of three distinct kinds of weakness of will. And the assumptions underpinning this set of explanations do not result in a puzzle. Recall from Section 3:

MSM THEORY	If, at time *t,* an agent's decision system D values some option A more highly than some option B, he believes he is free to do A, *and* PA allocates D for action-selection, then if he tries to do either at that time, he will try to do A and not B.

Adding...

WEAKNESS OF WILL	In cases like Gene's above, an agent judges that it is best to do A at *t*, believes he is free to do A at *t*, but, despite trying to do something, does not try to do A at *t*.

...does not constitute a puzzle. On the plausible assumption that judgment is underwritten by the deliberative system, weakness of will occurs in any case in which PA allocates either the hardwired or habitual systems for action selection.

The MSM solution has four consequences for relevant philosophical debates:

1. MSM can adjudicate between theories. As a computationally informed, empirically based theory of the mind, MSM provides independent grounds for preferring some theories over others.

MSM foremost rejects single system theories of the mind. Most philosophical theories of action assume a unitary progression of mental states connecting deliberation and action, where mental states proceed serially rather than in parallel (Davidson 1970, Bratman 1979, Audi 1979, Stocker 1979, Rorty 1980, Dunn 1987, Mele 1987, 1992, Buss 1997, Holton 1999). For instance, Rorty (1980, 334) describes the multiple "stages on thought's way to action," beginning with an agent's general considerations and ending with her acting according to her decision. But evidence from the decision sciences systematically contradicts the assumption of a single system model of the mind. It should be abandoned.

MSM aligns with the consensus view in the decision sciences to favor multi-system, interactionist theories. But MSM is more consistent with some multi-system views than others. While Plato famously discusses the tripartite soul in the *Republic*, his account describes *intra-systemic* competition between the faculties of a single, centralized soul or mind. MSM, on the other hand, describes *inter-systemic* competition between different systems operating in parallel. In doing so it provides partial support for two contemporary accounts of weakness of will: Davidson's (1982) partitioned mind hypothesis and Neil Levy's (2011) dual system account.

Davidson (1982) suggests that for weakness of will to be possible, one mental event, A, must cause another mental event, B, without A's being a *conclusive reason* for B. (If A were the conclusive reason for B, this would amount to a rational and not weak-willed action.) This indicates that there must be more than one path to action, i.e., more than just the path via conclusive reasons. Discussing how such a relation could be possible, Davidson proposes that weakness of will may require a partitioning of the mind.[35] Davidson's suggestion is supported by MSM: the mind consists of multiple, semi-autonomous decision systems.

MSM goes beyond Davidson's view, however, by providing a principle of arbitration to regulate interactions between systems. Since Davidson does not account for how parts

[35] Davidson writes: "If we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned into quasi-independent structures that interact ... Recall the analysis of akrasia. There I mentioned no partitioning of the mind because the analysis was at that point more descriptive than explanatory. But the way could be cleared for explanation if we were to suppose two semi-autonomous departments of the mind, one that finds a certain course of action to be, all things considered, best, and another that prompts another course of action. On each side, the side of sober judgment and the side of incontinent intent and action, there is a supporting structure of reasons, of interlocking beliefs, expectations, assumptions, attitudes and desires" (1982, 300).

interact, he leaves the exact workings of his view underspecified. Davidson recognizes: "we want to know why this double structure developed, how it accounts for the actions taken, and also, no doubt its psychic consequences and cure. What I stress here is that the partitioned mind leaves the field open to such further explanations" (1982, 300-301). MSM provides such explanations. Moreover, MSM demonstrates that the multiple interacting systems are profoundly advantageous to effective human choice, and not something to be 'cured' (see below).

MSM offers a weaker endorsement of Levy's (2011) theory of weakness of will. Levy presents an argument to identify weakness of will with dual process theory. Levy first identifies weakness of will with ego depletion, and then identifies ego depletion with interactions between System 1 and System 2.[36] Ego depletion occurs when acts of volition draw on limited resources, much like muscles use energy, and become used up (Baumeister *et al.* 1998, 2002). Levy proposes that weakness of will occurs when System 2 becomes depleted of energy, causing an automatic down-regulation to System 1. In some respects, the depletion view appears analogous to MSM. It identifies multiple systems and provides a (depletion-based) principle of arbitration to explain how they interact. It sets FOLK PSYCHOLOGICAL THEORY aside in favor of concepts from cognitive psychology. And it rightly aims to recast weakness of will in terms of a broader psychological phenomenon, namely, in terms of failures of self-control.

The depletion view faces a limitation that MSM does not, however. The depletion view explains a non-standard conception of weakness of will. The standard conception defines weakness of will as acting against one's better judgment (see Stroud 2003 for a review). But the depletion view explains acting against one's intention (proposed by Holton 1999). Levy further maintains, "our explanatory projects require us to abandon the notion of weakness of the will" (2011, 152-153). But since his view at best explains only intention-based weakness of will, the claim represents a substantial overstatement. By contrast, MSM explains and substantially develops the more widely held conception of weakness of will.

2. *Weakness of will must be fractionated.* MSM goes beyond introspection to identify three kinds of weakness of will. Each kind has a distinct computational mechanism, behavioral profile, and phenomenology. Fractionation is significant, because weakness of will is an everyday phenomenon that has become a theoretical construct. But if computational and empirical evidence suggests it is not a monolithic phenomenon, then the philosophical literature should instead be discussing a cluster of concepts.

[36] System 1 corresponds to unconscious, intuitive reasoning. System 2 corresponds to deliberate reasoning abilities (Kahneman, 2011).

Fractionation has additional implications for understanding degrees of knowledge and uncompelledness – or conversely, compulsion – in weakness of will. For instance, philosophers often debate what degree of knowledge is involved in weakness of will. While a Socratic *akrates* has merely some sense that she shouldn't be doing A (*Protagoras* 356c), a Davidsonian *akrates* is, in sharp contrast, explicitly and reflexively aware that she is acting against her own better judgment (1982, 180). But fractionation suggests that different types of weakness of will involve different degrees of knowledge. In pruning-based weakness of will, an agent cannot know that doing A would be preferable to doing B. She can only know that doing B will result in a negative outcome. But as her other alternative(s) have been pruned away, she does B. S only 'knows' what would be best in the sense that she knows it would be bad to do B, but does it anyway. By contrast, in inhibitory weakness of will, an agent fully represents two or more branches of the tree, and hence clearly knows that she should do A and not B. In this case, an agent exhibits so-called 'clear eyed' weakness of will, or what is analogous to unconditional weakness of will. Thus, we cannot make categorical statements about the degree of knowledge involved in weakness of will.

Fractionation similarly recasts the debate surrounding weakness of will and compulsion. Davidson proposed that weak-willed actions must be free and uncompelled (Davidson 1970). But others have argued for a more ambiguous relationship between weakness of will and compulsion (Watson 1977, Audi 1979, Mele 1987, Mele 2002). For example, Gary Watson (1977) argues that weakness of will occurs when an agent gives in to temptation but, had her regular capacities for self-control been developed, she could have resisted giving in. Again, fractionation suggests that different types of weakness of will involve different degrees of compulsion or lack thereof. The role of learning and updating based on experience in habitual weakness of will suggests that regular efforts to avoid such habits would indeed enable an agent to pursue the most valuable alternative. By contrast, even if an agent *believes* that she is free, the automatic, inflexible nature of the hard-wired system suggests that no amount of self-control can help her avoid either pruning-based or inhibitory weaknesses of will. Across a step-wise distribution, pruning-based and inhibitory weaknesses of will are closer to compulsion than is habit-based weakness of will.[37]

3. Weakness of will as a byproduct of choice. Folk psychological theories propose that an agent reasons rationally until something unexpectedly goes wrong. MSM weakness of will is the inevitable outcome of the regular workings of the machine. The multiple decision systems operate in parallel, trading off to optimize decision-making. In turn, these tradeoffs occasionally result in instances of weakness of will. Weakness of will is thus not a

[37] One question that follows from this, suggested by an anonymous referee, is what MSM entails regarding weakness of will and accountability. One implication of MSM, perhaps contrary to previous opinion, is that weakness of will is a relatively common phenomenon, so that the issue of how we should hold people accountable is of greater importance on the MSM worldview. It is plausible that, just as MSM fractionates different types of weakness of will and different degrees of compulsion, so it is consistent with a step-wise attribution of accountability. However, MSM does not commit us to any one view.

breakdown in the system or, as George Ainslie (2001) describes it, a breakdown of the will. It is a simple byproduct of everyday decision-making.

4. *Weakness of will as a blueprint for future research.* More broadly, due to its domain-general approach, MSM offers a promising theory of practical reasoning. In identifying the underlying causes of weakness of will, MSM captures a small part of what is likely a large and complex map of decision-making 'fault lines.' Understanding the parameters and interactions of our multiple decision systems enables us to identify more general patterns in decision-making. Understanding these patterns in turn provides the foundations for a comprehensive set of principles to guide our practical deliberations. The MSM discussion of weakness of will provides a blueprint for this future research program.[38]

References

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

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

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

Arpaly, N. (2000). On acting rationally against one's better judgment. *Ethics* 110: 488-513.

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

Austin, J.A. (1956/57). A plea for excuses. In Austin (1979), 175-204.

Austin, J.A. (1979). *Philosophical papers*, 3rd ed., J. O. Urmson and G. J. Warnock (eds.), Oxford: Oxford University Press

[38] Fault line approaches to the decision systems are gaining substantial ground in psychology and neuroscience. For example, neuroscientist A. David Redish uses the term 'vulnerabilities' to highlight those aspects of our decision-making architecture that are susceptible to illnesses such as addiction and depression. One key vulnerability consists in the mammalian opioid system and its susceptibility to external chemicals such as opium and heroin (Redish 2013). Quentin Huys and colleagues have similarly argued that psychiatric illnesses such as depression and impulsivity might be byproducts of our decision-making systems (2012, 2013). A similar approach should be taken up in the philosophy of action.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.

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

Balleine, B.W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37, 407–419 (1998).

Balleine, B.W. & Dickinson, A. The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. J. Neurosci. 20, 8954–8964 (2000).

Balleine, B.W., Daw, N., O'Doherty, J.P. (2009). Multiple forms of value learning and the function of dopamine. In Glimcher, P. W., Fehr, E., Camerer, C., & Poldrack, R. A., Eds. (2009), 367-388.

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.

Barrett LF (2006). Are emotions natural kinds? *Perspectives on Psychological Science* 1(1):28–58.

Baumeister, R.F., Bratslavsky, E., Muraven, M. and Tice, D.M. (1998). Ego-depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252–1265.

Baumeister, R.F. (2002). Ego Depletion and Self-Control Failure: An Energy Model of the Self's Executive Function". *Self and Identity* 1: 129–136.
Bobonich, C., and Destrée, P. (eds.), 2007, Akrasia *in Greek Philosophy: From Socrates to Plotinus*, Leiden, Boston: Brill.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Botvinick, M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

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.

Brown, P. L., & Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the experimental analysis of behavior*, *11*(1), 1-8.

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

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated?. *Frontiers in psychology*, *5*, 823.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: recent developments and current challenges. *Philosophical Transactions: Biological Sciences*, 1515-1527.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

Davidson, D. (1970). How is weakness of the will possible?. In Davidson (1980), 21-42.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1982). Paradoxes of Irrationality, in Davidson (2004), 169-187.

Davidson, (2004). *Problems of rationality*. Oxford: Clarendon Press.

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*(7095), 876-879.

Daw, N. D., & O'Doherty, J. P. (2013). Multiple systems for value learning. *Neuroeconomics: Decision Making, and the Brain*, 393-410.

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

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., & Huys, Q. J. (2008). Serotonin, inhibition, and negative mood. *PLoS Computational Biology*, *4*(2), e4.

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

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

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

Dennett, D. C. (1989). *The intentional stance*. MIT 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., (1994). Instrumental conditioning. *Animal Learning and Cognition,* Ed. N. Mackintosh. San Diego: Academic Press, 45-79.

Dickinson , A., Squire, S., Varga, Z., and Smith, J.W. (1998). Omission learning after instrumental pre-training . *Quarterly Journal of Experimental Psychology* 51, 271-286.

Dickinson, A. & Balleine, B. (2002). The role of learning in motivation. *Stevens' handbook of experimental psychology Vol. 3: Learning, motivation, and emotion 3rd Ed.,* Ed. C.R. Gallistel. New York: Wiley, 497-533.

Dunn, R. (1987). *The possibility of weakness of will*. Indianapolis: Hackett.

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

Glimcher, P. W. (2010). *Foundations of neuroeconomic analysis*. Oxford University Press.

Glimcher, P. W., & Fehr, E. (Eds.). (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Gosling, J. (2002). *The Weakness of the Will*. Routledge.

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.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Angonno, C. O., Batailler, C., Birt, A., et al. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science.*

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.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. Science, 324(5927), 646-648.
Heil, J. (1989). Minds divided. *Mind*, 571-583.

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, *14*(4), 443-451.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Holcombe, A. O. (2016). Introduction to a Registered Replication Report on Ego Depletion. *Perspectives on Psychological Science*, *11*(4), 545-545.

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.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

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., 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).

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of mood & anxiety disorders*, *3*(1), 1.

Izquierdo, A., Suda, R.K. & Murray, E.A. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J. Neurosci. 24, 7540–7548 (2004).

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: consensus and controversy. *Neuron*, *63*(6), 733-745.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. *Punishment and aversive behavior*, 279-296.

Kelley, A. E., & Berridge, K. C. (2002). The neuroscience of natural rewards: relevance to addictive drugs. *The Journal of Neuroscience*, *22*(9), 3306-3311.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759-764.

Killcross, S. & Coutureau, E. Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb. Cortex 13, 400–408 (2003).

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.

Levy, N. (2011). Resisting 'weakness of the will'. *Philosophy and Phenomenological Research, 82*(1), 134-155.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, 50: 249–58; reprinted in Rosenthal 1994, pp. 204–10.

Macintosh, N.J. (1983) *Conditioning and associative learning*. Oxford: Oxford University Press.

Mele, A. (1987). *Irrationality*. New York: Oxford University Press.

Mele, A. (2010). Weakness of will and akrasia. *Philosophical studies*, *150*(3), 391-404.

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). 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.

Montague, R. (2007). *Your brain is (almost) perfect: How we make decisions*. Plume Book.

O'Doherty, J. P. (2014). The problem with value. *Neuroscience & Biobehavioral Reviews*, 43, 259-268.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223-226.

Plato. (1997). *Complete Works*, J.M. Cooper, D.S. Hutchison (eds.). Indianapolis: Hackett, 746-791.

Pears, D. (1984). *Motivated Irrationality*, Oxford: Clarendon Press.

Peijnenburg, J. (2000). Akrasia, dispositions and degrees. *Erkenntnis*, *53*(3), 285-308.

Penner, T. (1997). Socrates on the strength of knowledge: Protagoras 351B-357E. *Archiv für Geschichte der Philosophie*, *79*(2), 117-149.

Rangel, A. (2013). Regulation of dietary choice by the decision-making circuitry. Nature neuroscience, 16(12), 1717-1724.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545-556.

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.

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 nonreinforcement. *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.

Santas, G. (1966). 'Plato's Protagoras and Explanations of Weakness of Will.' *Philosophical Review* 75: 3.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, *117*(2), 245-273.

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.

Simpson, M. (2003). *The metamorphoses of Ovid*. Univ of Massachusetts Press.

Sripada, C. (2010). Philosophical questions about the nature of willpower. *Philosophy Compass*, *5*(9), 793-805.

Spinoza, B. (2002). *Complete Works,* Samuel Shirley (Trans.). Indianapolis: Hackett.

Stich, S. and S. Nichols. (2003). Folk Psycholgy. *The Blackwell Guide to Philosophy of Mind*, S. Stich and T. Warfield (eds.), Oxford: Blackwell, pp. 235–55

Stocker, M. (1979). Desiring the Bad: An Essay in Moral Psychology. *Journal of Philosophy*, 76: 738-753.

Stroud, S. (2003). Weakness of will and practical judgment. In Stroud and Tappolet (2003), 121-146.

Stroud, S., and Tappolet, C. (Eds.). (2003). *Weakness of will and practical irrationality*. Oxford: Clarendon Press.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge: 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.

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.

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2012). Hierarchies in action and motor control. *Journal of Cognitive Neuroscience*, *24*(5), 1077–1086.

Uithol, S., Burnston, D. C., & Haselager, P. (2014). Why we may not find intentions in the brain. *Neuropsychologia*, *56*, 129-139.

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.

Velleman, D. (2000). *The possibility of practical reason*. Oxford: Oxford University Press.

Watson, G. (1977). Skepticism About Weakness of Will. *Philosophical Review* 86: 316-339.

Williams, D. R., & Williams, H. (1969). Auto-maintenance in the pigeon: sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior 12*(4), 511-520.