The Dissertation Committee for Amanda Lea Evans Certifies that this is the approved version of the following Dissertation:

Loss of Control and Phenomenology in Mental Disorder

Committee:

Michelle Montague, Supervisor

Galen Strawson

Ernest David Sosa

Tim Bayne

Hanna Pickard

Loss of Control and Phenomenology in Mental Disorder

by

Amanda Lea Evans

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2021

Acknowledgements

There are more people than I could reasonably include in an acknowledgments section who have helped me in some important way or other in getting to this point in my academic career. First and foremost, I am immensely grateful for my supervisor, Michelle Montague, who humored my energetic but often not-fully-formed interjections and philosophical ideas during my first couple years of graduate school and then took me on as her student. I will forever be grateful that I was given the chance to evolve philosophically and come into my own at UT under Michelle's, Galen Strawson's and David Sosa's kind and illuminating guidance. I am also very grateful for my two external committee members, Tim Bayne and Hanna Pickard, who have been invaluable in helping me develop philosophically as well as professionally in the past couple of years.

I would also like to thank my circle of confidants at UT who have stuck with me through thick and thin as I battled my own mental health issues and figured things out during my early and mid-twenties. Alicia Armijo, Emma McDonald, Whitney Benson, Stella Fillmore-Patrick, Anugya Sood, Karim Nader, Savanna Shaffer, Gabi Hitel, Brigitte Gill, Laurenz Casser, Brian Pollex, Hannah Trees, Henry Schiller, and Jake Galgon are the names at the forefront of my mind when I think of who has helped me grow, learn, and have fun during the past five years in Austin. In varying ways these people made up my Austin "family", and I am very grateful to have had them along for the ride with me in addition to my family and friends back home as well as my partner.

Lastly, I would like to thank the few people at Notre Dame who saw something in me and took a chance on me when I was, on paper, a bit of a risk to vouch for during the darkest years of my life. Had it not been for the advocacy of my advising dean, Collin Meissner, the readmission committee following my medical leave may not have been

persuaded to take a chance on readmitting a suicidally depressed anorexic. And, had Fritz Warfield, Peter van Inwagen, and Leopold Stubenberg not seen something in me philosophically, I am certain I would not have had the opportunity to pursue philosophy at the doctoral level. The irony of a former anorexic writing a dissertation on how anorexics are both obsessed with and not very good at analyzing the true contents of their inner lives is not lost on me, and I have everyone listed above to thank for helping me get to a place where I can pursue these philosophical projects that are so near and dear to my heart.

Abstract

Loss of Control and Phenomenology in Mental Disorder

Amanda Evans, PhD

The University of Texas at Austin, 2021

Supervisor: Michelle Montague

Any insights we can hope to gain with respect to what is going on with our mental lives and our agency will almost certainly require a close examination of the "worst-case scenarios", since it is when things break down that the joints of the phenomena are revealed. This is a philosophical intuition of mine that pervades everything I work on, and the papers that make up this dissertation are no exception. In keeping with this guiding sentiment, this dissertation tackles three philosophical issues related to the so-called "loss of control" that occurs in mental disorder, and it does so in a way that places the *phenomenology of agency* at the forefront in some way or other.

In my first paper on the sense of agency in anorexia nervosa (AN), I try to resolve an apparent discrepancy between the phenomenology of anorexics in the grip of their disorder and the psychological and neurological data that purport to describe what they are undergoing. I provide a solution to this apparent incongruency by offering an account of the sense of agency in AN that grants sincerity to anorexic testimony while also being able to explain *why* the relevant experiences of agency come to be illusory. Then, in my second paper, I broaden my scope to include not just AN but also substance use disorder (SUD).

After outlining the debate surrounding the question of whether addiction ought to be

categorized as a form of akrasia, I show that the phenomenon at issue is far more complex than either side has supposed. I then propose a "horseshoe model" of loss of control that is able to capture the complexity that is brought in by examining the similarities and differences between SUD and AN.

Finally, in my third paper, I pursue a question that arises from the exposition of the horseshoe model introduced in the previous paper. The question is, roughly, "Why is one 'half' of the horseshoe model associated with the phenomenology of loss of control while the other "half" is associated with the phenomenology of extreme self-control?". This line of inquiry ultimately leads to an understanding of how one's pathological desires can be experienced quite differently depending on the content of one's self-image. Taken together, it is my hope that these papers can contribute to the philosophical goal of unearthing the realities of our mental lives and our agency by examining the fault lines formed by psychopathology.

Table of Contents

List of Tables

List of Figures

INTRODUCTION TO THE DISSERTATION

Although this dissertation follows the so-called "MIT style" in that it is composed of three separate papers, the following articles are united in that they have been driven by the same philosophical skepticisms and intuitions that pervade just about everything I have worked on thus far in my philosophical career. These skepticisms and intuitions can be summed up by two (admittedly hyperbolic) claims: i.) I am deeply skeptical that we can know much of anything about our lives as agents with any amount of certainty, and ii.) Any insights we can hope to gain with respect to what is going on with our mental lives and our agency will almost certainly require a close examination of the "worst-case scenarios", since it is when things break down that the joints of the phenomena are revealed.

In this way, this dissertation can be seen as a three-pronged approach to beginning the research project suggested by this latter claim while the former claim lurks in the theoretical background. In particular, I have chosen to tackle three philosophical issues that relate to the so-called "loss of control" that occurs in mental disorder in a way that places the *phenomenology of agency* at the forefront in some way or other. My reason for this is very reminiscent of some of the main points made by Owen Flanagan in his chapter, "What is it like to be an addict?"[1]. In it, Flanagan, himself a former addict, stresses that a thorough and accurate understanding of addiction cannot discount the important data to be gained by inquiring into what it is like to be a *token* of the addict *type*. In other words, we as

[1] Flanagan, Owen (2011). "What is it like to be an addict?", in Poland, Jeffrey, and Graham, George (eds.), *Addiction and Responsibility*. Oxford University Press: Oxford.

theorists must not overlook *what it is like* to be an individual experiencing the mental disorders we theorize about. In keeping with this sentiment, I do not think we can adequately theorize about any of the various mental disorders without trying in earnest to unite the phenomenology of the token individual with what psychology and neuroscience have to say about mental disorder *x*.

My first paper, "Anorexia Nervosa: Illusion in the Sense of Agency", takes this sentiment in stride in that it seeks to resolve an apparent discrepancy between the phenomenology of anorexics in the grip of their disorder and the psychological and neurological data that purport to describe what they are undergoing. I provide a solution to this apparent incongruency by offering an account of the sense of agency in anorexia nervosa (AN) that grants sincerity to anorexic testimony while also being able to explain *why* the relevant experiences of agency come to be illusory.

Then, in my second paper, "A Horseshoe Model of Pathological Loss of Control", I broaden my scope to include not just AN but also substance use disorder (SUD). In this paper I begin by examining the debate surrounding the question of whether addiction ought to be categorized as a type of akrasia. I weigh in on this debate by first outlining a number of empirical factors that unite AN and SUD while also highlighting some key differences between the two disorders. With this empirical data in hand, I propose a "horseshoe model" of loss of control that places AN on one extreme end of the horseshoe and SUD on the other, thereby accounting for the deep similarities between the two conditions while also noting their differences.

Finally, in my third paper, "Alienation and Identification in Pathological Loss of Control", I pursue an interesting question that arises from the horseshoe model proposed in my second paper. This question is, roughly, "Why is one 'half' of the horseshoe model associated with the phenomenology of loss of control while the other "half" is associated with the phenomenology of extreme self-control?". In providing an answer to this question, I divert slightly from the empirically heavy methodology employed in the first two papers, and I instead harken back to a discussion of Harry Frankfurt's unwilling addict and his concept of externality. This line of inquiry ultimately leads to an understanding of how one's pathological desires can be experienced quite differently depending on the content of one's self-image. By approaching the phenomenology of loss of control from a slightly different angle, my third paper can be seen as an account that adds to and runs alongside the models that are argued for in the first two papers.

In closing, it is my hope that the following three papers contribute to the philosophical goal of unearthing the realities of our mental lives and our agency by examining the fault lines formed by psychopathology. It should be noted, however, that I do not think the conclusions reached in this dissertation to apply solely to agents with mental disorders. Rather, the accounts developed here should be seen as applying to *all* agents, albeit to lesser and varying degrees. The mentally ill do not have minds or powers of agency that are fundamentally different from the minds of those philosophers who have developed the current mainstream theories on mind and action that draw on their own phenomenological experiences and intuitions. And, indeed, there are many of us in philosophy who inhabit the worlds of mental illness and philosophy simultaneously. In

critiquing the tendency to "other" mental illness, historian Roy Porter once wrote that "[s]etting the [mentally] sick apart sustains the fantasy that we are whole"[2]. I believe the same sentiment is true when it comes to how we ought to theorize about the mind and agency—trying to "set the sick apart" can only result in philosophical theories that are part fantasy.

[2] Porter, Roy (2002). *Madness: A Brief History*. Oxford University Press: Oxford, p. 62-63.

ANOREXIA NERVOSA: ILLUSION IN THE SENSE OF AGENCY

It is, at the most basic level, a bundle of deadly contradictions: a desire for power that strips you of all power. A gesture of strength that divests you of all strength.

(Marya Hornbacher, Wasted: A Memoir of Anorexia and Bulimia)

0. Introduction

What is it like to live with anorexia nervosa? While it is doubtful that those without a history of the disorder can ever fully grasp what the experience of it entails, memoirs and other written works by individuals who have lived with anorexia nervosa (AN) can provide some insight. Reading through works such as Hornbacher's (quoted in the epigraph), Kelsey Osgood (2013), and Emma Woolf (2013)—Virginia's great niece, who claims Virginia herself was anorexic—one quickly notices recurring themes that weave their way throughout the various autobiographical accounts. One of these is an intense fixation with concepts that philosophers tend to be similarly interested in— musings on self-control, willpower, and ambivalence in acting are standard fare in first personal accounts of anorexia nervosa[4].

Although there is no doubt wide variation in the lived experiences of those diagnosed with AN, in sifting through published autobiographical works as well as data from qualitative studies two generalizations present themselves as apt. First, anorexics[3]

[3] There has been recent discussion in certain areas of literature regarding the exclusive use of person-centered language (in this case, "individual with anorexia nervosa") in lieu of traditional descriptors for individuals with mental disorders (here, "anorexic"). A proper treatment of my views on this issue would take me beyond the scope of this paper, although for present purposes I will note that while I do agree that person-centered language can be useful in certain contexts, I have theoretical as well as practical reasons for thinking the term "anorexic" should continue to be used alongside it depending on the context. One such reason is that anorexics commonly refer to themselves as "anorexic", and this paper focuses

tend to be much more concerned with honing and maintaining their powers of agency than the average person. Second, the gradual deterioration into the disordered state of anorexia nervosa, if we are to take seriously the past several decades of research on the disorder, ultimately results in a serious curtailment of some of the same agentive capacities that were prized by the anorexic individual at the outset. This peculiar situation that severely ill anorexics find themselves in vis-à-vis their apparent lack of agency appears most often in philosophical literature in the context of applied ethical dilemmas concerning personal autonomy and compulsory treatment (Cf. Draper 2000,
Giordano 2005).

However, the bioethical debate that hinges in part on the actual status of the anorexic's powers of agency is not the focus of this paper. Rather, the present account seeks to resolve the "contradiction" of anorexia nervosa that Hornbacher alludes to. The first stated half of this contradiction—the "desire for power"—meshes well with first personal reports of what it is like to live with anorexia nervosa in the early stages of illness. On the other hand, the second component—the stripping of power—coheres well with the current empirical understanding of anorexic food restriction *as well as* the reports of anorexic patients who have sufficiently progressed in the recovery process.

In this paper I will first show that the two accounts of anorexia nervosa we ought to take seriously—that is, the first personal reports of those who have experienced it firsthand as well as the research that seeks to explain anorexic behavior from an empirical

on the lived experience of the anorexic individual *qua* anorexic. For ease of exposition, then, I will continue to use the term "anorexic" alongside the phrase "individuals with anorexia nervosa".

perspective—appear to be thoroughly in tension with one another in their descriptions of anorexic actions. Rather than proceeding at this point by way of disregarding anorexic testimony as meaningless or insincere, I will instead offer a positive account of the sense of agency in anorexia nervosa that renders these two depictions compatible. The resultant picture of anorexic behavior is one that accommodates current empirical findings while also providing valuable insight into how it is that anorexics can sincerely report feeling fully in control over their food restriction.

The paper will proceed as follows. In §1 I will introduce anorexia nervosa from the perspective of anorexics' reports. Then, in §2, I will discuss empirical theories that aim to explain the development and persistence of AN while also noting the ways in which these accounts are in tension with those discussed in §1. Finally, in §3, I will resolve the tension between the descriptions of AN discussed in §1-2 by offering a positive account of how the sense of agency in anorexia nervosa comes to be illusory.

1. Anorexia nervosa from the anorexic's perspective

In order to appreciate the ways in which anorexics might be mistaken about the nature of their condition one must first have a sense of what, exactly, the common experiences are amongst anorexics that might be inaccurate. The aim of this section is to provide such a gloss, although it must be stressed that experiences do, of course, vary greatly across the anorexic population. That being said, it is indisputable that there is a shared body of experiences and conceptualizations that many individuals with anorexia nervosa share. Megan Warin (2004), who interviewed and got to know forty-six anorexic

participants over a 15-month period while conducting an ethnographic research study, described it thus:

> *Like many people who share a common diagnosis, those with anorexia shared an understanding of the symbolic power of anorexia, and the contradictory desire to be the thinnest, the sickest and therefore the most successful… Collectively, participants referred to 'the secret language of eating disorders', a language that was articulated through a range of body practices and knowledges, such as… the proudness associated with the 'hard, clean truth' of jutting bones (p. 101, emphasis mine).*

These experiences, which are outlined by Warin and others in qualitative, interview-based studies, provide valuable phenomenological data that I will go on to argue is in tension with the scientific understanding of anorexic behaviors.

Even with the help of qualitative studies, however, typifying the anorexic experience is complicated by the fact that it tends to progress in stages (Cf. Osler 2020, Warin 2004). There are two stages of AN that are relevant to the present account, and for the sake of simplicity I will be referring to them as the "pre-awareness" and "post-awareness" stages. The basic idea behind this bifurcation is to separate the times during which anorexics sincerely feel and believe that they are in control over their food restriction and the times after which anorexics have come to realize they are not in control. The latter category of experiences will be covered in §3, but for present purposes I will be discussing only the former category, which is sometimes referred to as the "honeymoon phase" of anorexia. While the transition from pre-awareness to post-awareness stage most often occurs in the context of clinical intervention, this will not always be the case for each

individual[4]. Furthermore, there is no requirement that all individuals reach the post-awareness stage, just as there is no requirement on a theory of substance abuse that all individuals will progress through a certain stage of acceptance or pursue recovery.

To that end, the following two excerpts are from anorexic participants who were asked to recount what it was like for them during the pre-awareness stage of AN:

> *I felt I was in better mood when I didn't eat. I had control, was on top of the situation. I compared myself to other people and then I felt privileged that I could control myself when tempted to eat (participant quoted in Nordbo 2006, p. 560).*

> *You know you feel very, very, calm and comfortable and sort of I guess safe, a mixture of all those sorts of things. And sort of security and sort of just*
>
> *RIGHTEOUSNESS as if this is the right thing… it's a very nice way to feel (participant quoted in Charland et al. 2013, p. 357)*

And, recounting the common sentiments expressed by her interviewees, Warin stated,

> *[A]norexia was, most particularly in its early phases, experienced as a productive and empowering state of distinction—some even referring to this stage as 'the honeymoon phase'. Others were eager to be diagnosed with anorexia as it was not experienced as a debilitating illness, rather, it was 'unique', 'heroic', 'an achievement' and 'a thrill' (Warin p. 101, emphasis mine)*

What these first-personal reports convey is that AN, at least in the pre-awareness stage of illness, tends to be experienced as a positively-valenced project emblematic of self-control as well as a source of pride and meaning. In the following section, I will provide an

[4] By adopting this rough categorization, I do *not* intend to suggest that there is a single point at which anorexics come to realize the nature of their condition. Although this may, of course, happen for some (relatively lucky) individuals, most often the process of becoming aware of and of processing one's situation is more of a back and forth process that can take months or even years. And, much like substance use disorder, achieving acceptance that one is no longer in control over one's behavior is only one of many sufficient conditions required for achieving some level of recovery.

overview of the current empirical understanding of these same behaviors that are experienced by these individuals as "righteous" indicators of self-control and of achieving one's goals.

2. Anorexia nervosa from the clinician's perspective

Sincere as the reports of anorexics in the pre-awareness stages of their disorder may be, they are fundamentally at odds with the claims made about AN by clinicians and empirical researchers. Indeed, if the experiences of engaging willfully in food restriction were entirely veridical, anorexia's status as a mental disorder in need of intervention and treatment would be dubious. Fortunately, this conflict is not merely a matter of patient testimony versus that of mental health professionals, due to the fact that individuals who recover from AN report realizing that they were mistaken about the nature of their condition—more on that later. This fact in and of itself suggests the existence of a puzzle regarding self-awareness in anorexia that is pre-theoretic insofar as it does not depend on the vindication of any particular empirical theory regarding the nature of AN. The phenomenological data of anorexics before and after gaining insight into their conditions suggests that in the former stage of illness anorexics are mistaken in *some* meaningful way about their condition.

In this section we will cover what, exactly, is pathological about food restriction in anorexia nervosa according to current psychological and neurological accounts. As we shall see, however, these theories can only explain the mechanisms by which pre-anorexic behaviors become relevantly pathological and are thus sustained. They do not offer any

explanation as to *why* the anorexic subject herself fails to recognize this transition into pathological behavior, which will be the task of the following section. First, though, we must consider the empirical research on AN in order to appreciate the substantial tension between the clinical-theoretic descriptions of anorexic behaviors and anorexics' experiences of these same actions. The research in question involves two theories for the pathogenesis and persistence of AN that have become increasingly popular. Although the researchers working on these theories tend to consider the two models to be compatible, they nonetheless differ in terms of emphasis[5].

The first theory, which I will refer to as the "Habit Model" of anorexia nervosa, attempts to explain why anorexia nervosa has proven so difficult to treat when compared to other eating disorders. The case for the Habit Model is articulated most clearly in Walsh (2013). Walsh is concerned with arriving at a better understanding of what he calls anorexia nervosa's "enigmatic persistence", referring to the fact that AN as a disorder has remained markedly refractory to treatment despite significant empirical study and attempts to develop more effective treatment methodologies. He cites, for example, Steinhausen's (2002) findings that indicate that the outcome for anorexia nervosa, particularly for adult sufferers, did not improve substantially during the second half of the twentieth century.

[5] Note, however, that the present account does not rely on the two models being compatible. If it is revealed that only the Reward Model (or only the Habit Model) is an accurate account of anorexic pathogenesis, the underlying nature of the disordered actions that make up the anorexic condition will still be other than what the anorexic herself experiences, which is all that is required for my account. The ultimate purpose here is to show that the empirical literature and the anorexic's phenomenological testimony can ultimately be rendered compatible.

Although treatments such as cognitive-behavioral therapy (CBT) and selective serotonin reuptake inhibitors (SSRI's) are known to be effective for treating related disorders such as bulimia nervosa, mood disorders, and anxiety disorders, they are surprisingly ineffective at treating AN (Attia 2010). What *is* known about treating AN is that adolescent patients and those with a relatively short duration of illness are significantly more likely to achieve remission, whereas adult sufferers (even those in their 20s) and those with a longer duration of illness have poorer treatment outcomes and high relapse rates (Kaplan et al. 2009).

Walsh suggests that an explanation for this marked difference in treatment outcomes can be found in the neural mechanisms that underlie habit formation. He proposes that by the time an individual develops full-blown anorexia her dieting behavior has become encoded as habit, as opposed to being the result of ordinary, purposeful dieting actions. In Walsh's own words,

> *[T]he dieting behaviors of individuals with anorexia nervosa begin as goal directed actions that lead to weight loss, which is [experienced as] highly rewarding (action-outcome learning). Over time, the dieting behaviors are engaged in persistently and repeatedly and thereby become overtrained and habitual (stimulus-response learning) (p. 479).*

Here, Walsh is employing a theory of habitual action in which an action is labeled as "habitual" when it meets the following criteria: it is not innate, it is engaged in repeatedly, and it is not the result of conscious, sustained effort. It is important to note, however, that under this relatively minimal description of habit one can still be in control of and aware of one's behavior while performing the relevant action—in other words, this is not meant to be something akin to an automatic reflex. Rather, a habit in this sense is meant to be a

behavior that becomes increasingly over-selected and that requires less effort and planning to initiate than a non-habitual action.

In articulating the Habit Model, Walsh begins with the datum that dieting behavior is highly prevalent within Western cultures, particularly among young women and adolescent girls. However, most of these dieters do not go on to develop anorexia nervosa. Those individuals who *do* become anorexic will begin with typical dieting behavior but will at some point "cross over" into behavior that more closely parallels stimulus-response behavior. Stimulus-response conditioning involves an acquisition of a non-innate behavior (e.g. extreme dieting) that is relatively insensitive to the receipt of the initial reward once it has been well-learned. At this point, according to Walsh, the anorexic individual's dieting becomes so overtrained that it ceases to be merely instrumental to the reward of weight loss.

The result of this process is that the dieting behavior *itself* becomes encoded as habit in anorexic individuals. The anorexic begins her weight loss endeavors at the level of goal-directed action-outcome learning, which she finds substantial success with and experiences as highly rewarding. What ultimately sets the anorexic apart from her "normal" dieting peers, however, is that at some point in time the dieting behavior becomes intrinsically rewarding to the anorexic. Indeed, the setting under which anorexia nervosa typically develops makes it exceedingly likely that anorexic behaviors will become encoded as deeply entrenched habits, as opposed to the relatively innocuous everyday habits that tend to be easier to control. For one thing, eating disorders typically develop during a period of stress, and behaviors acquired during periods of stress are especially

prone to becoming habitually encoded (Schwabe and Wolf 2009). Furthermore, one of the primary findings from the infamous Minnesota starvation study conducted during World War II is that significant weight loss tends to increase compulsive patterns of behavior (Keys 1950). The result in the anorexic case is a vicious cycle of weight loss and habit reinforcement. In support of this connection, weight gain in anorexic patients is associated with decreased levels of obsessionality (Olatunji et al. 2010).

Since it was first proposed in 2013, Walsh's theory has garnered further empirical support. In one recent study, Coniglio et al. (2017) found that measuring the strength of habitual food restriction in anorexics was a better predictor of actual food restriction than measures of "effortful, goal-directed restraint" (p. 146), and concluded that their "findings support Walsh's hypothesis that food restriction is maintained through habitual, rather than goal-directed behavior in both individuals with AN and atypical AN" (p. 147). Furthermore, Steinglass et al. (2018) found that "targeting habit strength yielded improvements in clinically meaningful measures" in comparison to standard psychotherapy in a study of anorexic participants, which led them to conclude that "[t]hese findings support a habit-based model of AN, and suggest habit strength as a mechanism-based target for intervention" (p. 2584).

Despite this, the Habit Model is not the only game in town. A related yet distinct theory which I will refer to as the Reward Model proposes that AN develops through a process that closely mirrors the development of addiction according to Robinson's and Berridge's (1993) incentive sensitization theory. Although a thorough discussion of this theory and its application to anorexia nervosa would take me beyond the scope of the

present proposal, I will briefly note that disorder-specific cues in AN such as photos of emaciated and exercising bodies have been theorized to play a similar role to that of disorder-specific cues in substance abuse (Park et al. 2014, O'Hara 2015). According to Robinson and Berridge, in addiction the dopaminergic (i.e. reward) system becomes overly sensitized to drug-specific cues, which in turn leads to drug-seeking and drugtaking behaviors becoming increasingly compulsive in nature[6]. According to the Reward Model of anorexia, a similar process leads to anorexia-specific cues becoming increasingly sensitized and thus increasingly influential over anorexic behavior. Over time, the sensitization toward these disorder-specific cues contributes to the increasing compulsivity and rigidity of anorexic food restriction.

Both the Habit Model and the Reward Model appear to shed light on the fact that anorexics tend to find it extremely difficult to resume normal eating once they have committed to recovery. This is because both theories predict that simply deciding to commit to recovery is not sufficient, since what is really needed is behavioral intervention therapy aimed at disrupting the anorexic's habitual (or cue-driven) food restriction (Cf. Steinglass et al. 2018). This is consistent with the observations of one anorexic participant interviewed by Hope et al. (2013), who reported,

[6] A crucial element of Robinson and Berridge's theory is that incentive sensitization can lead to a decoupling of "wanting" and "liking" within the addict's dopaminergic reward system. As a result, addicts can seek out and "want" to continue taking drugs even when they do not straightforwardly "like" them. Although this may be applicable to individuals with AN who are in recovery but have not yet succeeded in ceasing anorexic behaviors, it is worth noting that such a decoupling is unlikely to occur within the mind of an anorexic who straightforwardly still "likes" the reward of weight loss and its associated effects.

> *Well I always THOUGHT that I could, like before I tried it I thought all the time well I could easily eat more and stop this if I wanted. But when I came to try to do that I couldn't (p. 24).*

If either or both of these models are correct in their assertions, however, they would account for one perplexing feature of AN (i.e., why it is so difficult for anorexics pursuing recovery to simply "eat more") while unwittingly introducing another. That is, if we are to take seriously the claim that purportedly anorexic actions are *not* the result of effortful restraint but of habitual or cue-driven behavior, then we must ask ourselves why this would appear to be at odds with the anorexic's own experience of her dieting behavior during the pre-awareness stage. Referring back to §1, recall that anorexics in fact tend to experience their food restriction as being the prime example of their willpower. However, both the Habit and Reward Models' descriptions of these same actions would predict an experience of acting that is quite unlike the experience of willfully accomplishing a goal. Resolving this tension will be the objective of the following section.

3. Reconciling the phenomenological and clinical descriptions of anorexic food restriction

Up until this point we have explored two different narratives pertaining to the development and maintenance of AN as a condition. According to one, anorexia nervosa is experienced by the subject as a willful and meaningful series of actions in pursuit of a goal. This is the phenomenological description of AN according to the subject, and it appears to conflict directly with the empirical theories of AN that draw from psychology and neuroscience. According to the Habit Model and the Reward Model of anorexia nervosa, food restriction in AN is triggered by either pathological habit formation,

incentive sensitization to disorder-relevant cues, or some combination thereof. If we wish to take seriously the sincere reports of anorexic individuals (both before and after recovery) as well as the current research that advocates for the Habit and Reward Models, an account is needed that enables us to interpret these two narratives in a way that is no longer incompatible.

A response that some may find prima facie plausible to the question of why there exists a discrepancy between the clinical and first-person phenomenological descriptions of anorexic food restriction is that it is due to the incidence of anosognosia in the anorexic population. Anosognosia, which translates from Greek as "ignorance of disease", is term that was originally used to describe stroke or brain injury victims who are unable to recognize that they have become paralyzed (Cutting 1978, Heilman 1991). In the context of anorexia nervosa, anosognosia is often used synonymously with "denial of illness". Although the distinction is not always made in the literature, however, a more precise description of anosognosia in the context of AN would be that it is the impaired self-awareness that *leads* to the denial of illness, rather than the denial itself (Cf. Vandereycken 2006). What I have been calling the "pre-awareness" stage of anorexia nervosa is, in effect, the stage during which the symptom of anosognosia will be most prevalent.

However, it is important to realize that "because she is anosognosic" is a tautological response to the question of "Why doesn't the anorexic individual accurately experience her food restriction as being pathologically habitual or cue-driven, as the research suggests?". This is because both *anosognosia* and *denial of illness* are merely descriptive terms in that they convey *that* anorexics appear to be missing something or

getting something wrong with respect to their condition. They are entirely silent as to the causal story of *how* this comes to be—in other words, they have nothing to say about the *why* question. In essence, responding "because she is anosognosic" to this question amounts to saying, "She is unaware of the nature of her actions because she lacks awareness with respect to the nature of her actions", which is clearly circular. For this reason, citing the symptom of anosognosia in this case is an explanatory nonstarter.

3.1 Toward a Solution: The Sense of Agency

Fortunately, we can do better than this tautological answer to our question, which can now be slightly reformulated as: "Why, if we are to accept researchers' claims that anorexic food restriction is pathologically habitual or cue-driven, do anorexics exhibit anosognosia with respect to the nature of these behaviors?" In other words, we are after a way to reconcile the fact that pre-awareness stage (i.e. anosognosic) anorexics experience their food restriction as effortfully performed as opposed to habitual (or unreflectively selected, as is the case with cue-driven cravings).

In order to accomplish this, however, we will require some technical machinery that has the ability to describe veridical and falsidical experiences of one's agency, since this is the phenomenon that requires explication in the anorexic case. To do this, I will adopt a framework that has been developed in the literature on the sense of agency, which is the interdisciplinary subfield that seeks to explain the structures that underpin our phenomenologies and judgments of our actions. Following Tim Bayne and Elisabeth Pacherie (2007) I will speak of agentive phenomenology (i.e., the raw phenomenological

feel of performing an action) as separable from agentive judgments (i.e., the judgments associated with a given action). Furthermore, I will use the term "sense of agency" interchangeably with the term "agentive awareness", both of which are umbrella terms for grouping agentive phenomenology and agentive judgments together.

What, exactly, is the sense of agency supposed to be? As a (very brief) introduction to what is meant by the sense of agency in a non-pathological context, I invite you to imagine what it is like to perform a strength training exercise with a particularly heavy weight or to make the final push toward the end of a long and tiring run (or whatever other challenging action you choose). In actions such as these, the *phenomenology* of agency is especially vivid: the urgently fatigued feeling of one's wobbling limbs as one tries to stay the course, the feeling of effort required to continue pushing one's legs forward, etc. It is also true that an agent may judge herself to be performing these physically exerting actions, but her sense of agency in these cases would be much richer than that.

So, in addition to judging that she is pushing herself toward the end of her exercise, and in addition to her proprioceptive feelings of fatigue, the idea underpinning the entirety of the sense of agency literature is that that there is something it is *like* for an agent to be willing and controlling these very actions. It is also worth noting that when we are sufficiently "in the zone" while exercising we need not be judging much of anything at all. In these cases, we can still have agentive phenomenology (and thus agentive awareness) in much the same way that we can have visual phenomenology without making any associated

visual judgments in our less attentive moments[7]. If one is convinced by vignettes such as these, then one accepts that there is a distinctive sort of phenomenology that is inextricably tied to acting—that the sense of agency is not merely a matter of post-hoc cognitive judgments regarding action (see also Bayne 2008, 2011 for a more thorough treatment of these types of motivating cases).

Apart from the project of *describing* the phenomena relevant to the sense of agency, the bulk of the literature is devoted to arguing for or against various models of how the sense of agency is generated and structured. The relevant models of the sense of agency can be divided into those that claim that the sense of agency is generated exclusively by high-level cognitive states, those that claim it is generated exclusively by low-level sensory states[8], and those that view these two approaches as complementary rather than as theoretical rivals. Once again adopting terminology from Bayne and Pacherie (2007), I will refer to the first category as the "narrator" approach to modeling agentive awareness and the second as the "comparator" approach.

[7] I am using "in the zone" to refer to instances in which the agent is hyper-focused on the action she is performing and trying to maintain control of in the face of physical fatigue. It seems in these cases that the cognitive states necessary for producing judgements need not also be present, and oftentimes will not be. There is another sense of "in the zone" most commonly associated with running in which a subject's agentive phenomenology may also be diminished. Since I only wish to claim that we *sometimes* experience rich agentive phenomenology without associated agentive judgments, this is not a problem for the present point. If one associates being "in the zone" with the sort of diminished phenomenology commonly associated with running, then one can think instead of the weight training case. From my own experience, the weight training analog of "in the zone" seems to fit well with what I am describing.

[8] For ease of exposition, I will not be distinguishing between sensory states and perceptual states since the relevant takeaways regarding the integrated model do not depend on this distinction. Note, however, that Pacherie (2008, 2010) does distinguish between high-level cognitive states, intermediate-level perceptual states, and low-level sensory states in articulating her own variation of the integrated model for agentive awareness.

Philosophers who analyze agentive awareness in terms of a so-called "narrator" module believe that the sense of agency is generated entirely by a holistic, central systems mechanism that is in the business of producing high-level states such as beliefs, intentions, and inferences (e.g., Mylopoulos 2014, 2017). Proponents of the narrator approach will view agentive awareness as resulting from the mind's attempts to maintain and develop narrative self-understanding, albeit at a subconscious level. Put simply, the narrator approach claims that the sense of agency is governed by top-down processes concerned with inferences to the best explanation and maintaining coherence with one's occurrent intentions. One will experience one's behavior and will produce agentive judgments in ways that make sense, rendering the sense of agency a sort of fallible interpreter mechanism.

In contrast to the high-level narrator approach, comparator accounts of agentive awareness claim that our sense of agency is produced by atomistic mechanisms in the brain that are primarily concerned with motor control. This approach gets its name from the comparator model of the sense of agency first developed by Chris Frith and colleagues (Frith 1992, Blakemore and Frith 2003). The basic idea as it relates to agentive awareness is that a subject will experience movements as self-generated so long as there is a sufficient degree of match between the expected consequences of a given movement and the actual sensory feedback deriving from said movement. If there is too high a degree of mismatch, however, the subject will experience the movement in question as having been externally (or involuntarily) generated. The potential for fallibility according to a comparator-only account would primarily be a matter of local dysfunction within the motor cortex.

Finally, there are those who opt to embrace both the narrator and comparator models of the sense of agency. In defense of this final category, Pacherie (2010) notes that there is "a growing consensus that these different models should be seen as complementary rather than as rivals and that the sense of agency relies on a multiplicity of cues coming from different sources" (p. 446). Similarly, Moore (2016) endorses what he calls a "cue integration theory" for the sense of agency in which agentive awareness is generated by a combination of sensorimotor cues as well as top-down inferences related to "apparent mental causation", which is his terminology for the interpretive narrator module.

Crucially, both Moore (2016) and Bayne and Pacherie (2007) explicitly state that the resultant structure of the sense of agency according to an integrated model has significant potential for abnormal functioning in so-called "disorders of agency". Most of this discussion tends to center around schizophrenia, which has been hypothesized to be at least partially caused by abnormal processing of sensorimotor cues. However, this is not the only case of potential dysfunction according to the structure of the integrated model. Bayne and Pacherie note that the integration of these two approaches makes it possible for the narrator module to interfere with or even "override" the low-level deliverances of the comparator system. This sort of narrative interference would affect the resultant phenomenology experienced by the subject, which would in turn affect the agentive judgment based on said phenomenology. Alternatively, they suggest that the outputs from the sensorimotor system are often fleeting and ambiguous, meaning the narrator module will often have to "fill in the gaps" in its interpretation of the information. Similarly, Moore suggests that the relative influence of the comparator system versus the "apparent mental

causation" (i.e. narrator) system may be influenced by their apparent reliability as well as other standing psychological factors. In other words, the structure of the integrated model allows for various forms of non-veridical contents making it into one's agentive awareness, depending on the interplay and weighting of the comparator and narrator systems.

3.2 Egosyntonicity and the Solution to the Puzzle

Bringing the focus back to anorexia, we have one final piece to add before we are finally in a position to answer the question of *why* anorexics do not experience their food restriction (during the anosognosic stage) as habitual or cue-driven. This final element is that anorexia nervosa is considered to be an *egosyntonic* disorder, meaning that its sufferers tend to identify with the goals and behaviors that are part and parcel of the disorder itself (O'Hara et al. 2015, Gregertsen et al. 2017). Indeed, one of the many reasons that anorexia nervosa is so difficult to treat is that many of its sufferers view their disorder as exemplifying the perfectionism and powers of self-denial that they take to be core elements of their identity and values (recall the excerpts quoted in §1 of the study participants who opined about how "great" and "righteous" they felt during the pre-awareness stages of their eating disorders). In keeping with this, Vitousek et al. (1998) write that "[t]he anorexic's behaviors of food restriction and exercise are fully consonant with her goals of thinness and self-control" (p. 392-393), and Warin (2004) noted how some of her study participants even described their anorexia as a "friend" or "lover" (p.101).

The fact that anorexia nervosa is egosyntonic in this manner makes it unique among the other mental disorders (Gregertsen et al. 2017). I will now show that it is also an

important explanatory component of the solution to the puzzle we started with. Recall that the answer we are after is some sort of explanation for why anorexics' sense of agency is such that they believe they are both willfully and effortfully engaging in food restriction when the science says otherwise. To anticipate, a rough formulation of my solution is that the anorexic's sense of agency does not reflect this because her pre-awareness stage food restriction is, in a sense, causally overdetermined.

In order to put some flesh on this proposal, recall from §2 that both the Habit Model and the Reward Model involve a sort of cross over from ordinary goal-directed dieting behavior to actions that are pathologically habitual or cue-driven in nature. Despite this, the goals and intentions of the anorexic remain unchanged even though the underlying causal basis of the behavior has changed on a neurological level. Humans cannot simply intuit a shift in the underlying neurological bases of their actions, however—they have to go off of observable evidence. And it is here that the integrated model of the sense of agency becomes salient, given that it is designed to offer insight into how we come to gain awareness and insight into our actions.

In applying the integrated model to the anorexic case, Moore's (2016) description of a "theory of apparent mental causation" for the narrator module is especially illuminating. The idea is that, absent any reason to conclude otherwise, the anorexic continues to believe that her goal-directed willpower is still the causal source of her food restriction that she takes to be effortful. And, without significant evidence to the contrary, this is indeed a rational inference on the part of the narrator module. Despite this, the Habit Model and Reward Model must say that although the anorexic's long-term goals and

willfulness were causally efficacious before the full-blown development of AN, it is at that point no longer the impetus for food restriction[9]. In a way, this causal overdetermination of sorts should come as no surprise, given that anorexia nervosa is so perplexing and unique *precisely because* the behaviors that otherwise bear striking resemblance to compulsive drug abuse happen to be the same types of actions that the individual was set on performing before the onset of pathological functioning.

Given that the integrated model involves both agentive phenomenology as well as agentive judgments, one might well wonder where, exactly, I am intending to locate the source of the falsidical contents of effortfulness within the anorexic's sense of agency. Unfortunately, I do not know of a straightforward way to exactly pinpoint this phenomenon of illusory willfulness within the structure of the sense of agency. In fact, this difficulty is arguably built into the highly integrated structure of the model itself. When describing the various ways that the narrator and comparator might interact, Bayne and Pacherie list i.) a case in which the comparator system generates some phenomenological contents that the narrator then dismisses, ii.) a case in which the narrator enacts some version of cognitive penetration to actually alter the contents of the agentive phenomenology produced by the comparator system, and finally iii.) a case in which the contents given by the comparator are either minimal or are "dampened" such that the narrator has considerable leeway in "filling in" the gaps with respect to the agent's agentive judgment. Without some way of having access to an intensively detailed report of anorexics' agentive phenomenology and

[9] Note that this proposal inherits a virtue of Walsh's (2013) Habit Model of the pathogenesis of AN in that it contains a non-ad hoc and theoretically meaningful point at which an individual can aptly be given the anorexic label.

agentive judgments throughout the day as they engage in disordered behavior, I must admit that I do not see any clear way to definitely decide among these possibilities at this level of specificity.

Despite this, I do not see this as a dealbreaker for the use of the integrated model in accounting for the puzzle. In fact, my suspicion is that a combination of the above possibilities is at play in the anorexic case, and that the relative frequency of these phenomena vary from person to person and even across time for a particular individual. In general, the agentive phenomenology associated with routine habitual actions is not usually particularly rich or striking—in contrast to the vignette of tiring exercise envisaged earlier on, the agentive experience of habitually brushing one's teeth is far duller. This bodes well for all three options, since the relatively weak and uninteresting contents that *should* be informing the anorexic's sense of agency would be much more like the case of the teeth brushing than the exercise and would therefore be relatively minimal[10].

Lastly, it is worth noting that extensive research has been done on the apparently diminished interoceptive capacities of anorexics. Anorexics exhibit deficiencies in interoceptive awareness of bodily states such as heartrate (Pollatos et al. 2008), and they also perform poorly on tasks that require proprioceptive integration (Case et al. 2012). Papezova et al. (2005) have also hypothesized that the elevated pain threshold noted in

[10] What I am calling "illusory willfulness" would amount to a mistaken agentive experience of effortfulness or else a mistaken agentive judgment to that effect. In locating the feeling or judgment of effortfulness within the sense of agency I am assuming that relatively rich contents are present in the sense of agency. This is very much in the spirit of Bayne and Pacherie's (2007) account, given that they believe a "strong case can be made" that "the degree to which an action is effortful" can be included in the contents of agentive awareness (p. 477).

anorexic populations is due to generally diminished interoceptive awareness, and Jacquemot and Park (2020) suggest that diminished interoceptive capacities are a contributor to body dysmorphia. Although I am not aware of any study that measures anorexic interoceptive capacities relating to sensorimotor action cues in particular, the apparently widespread deficiencies of interoception bode well for the present account. This is because the proposed outputs of the comparator model as described by Frith and colleagues are just the sort of low-level internal sensory states that have been implicated in diminished interoceptive awareness in AN.

To further illustrate in what sense the sense of agency in anorexia nervosa is illusory, note that we do not ordinarily experience our habitual actions as resulting from our conscious effort and values[11]. That is, we do not experience habitual actions as being the *direct* result of effort and willpower *as they are occurring*. We may put a great deal of effort into trying to develop habits we consider to be beneficial, but that is not analogous to the present case. The goal, after all, of trying to develop healthy habits is to reach a point wherein the exercise regimen or healthy eating becomes "second nature" and thus no longer requires significant willpower to perform in the moment. Ordinarily, then, habitual actions are not experienced as involving significant effort or will once they have already become encoded as habit. In the anorexic case, however, this is precisely what is occurring.

[11] Since actions that are cue-driven are not typically recognized as such, there would presumably not be much in the way of phenomenology of cue-driven actions that one could compare to effortful actions. That being said, whichever way they are experienced on the personal level, it is unlikely that cue-driven behaviors are experienced as the direct result of conscious effort and willfulness.

One point in favor of this proposal is that it accurately predicts the shifts in the anorexic experience that occur during the post-awareness stage wherein anosognosia is reduced. As has already been noted, the clash between the anorexic pre-awareness experience and the Habit and Reward Models is pre-theoretic in that it is already anticipated in the reports of individuals who are no longer in the pre-awareness stage. This is because, when anorexics begin to try doing the *opposite* of what they had been doing (i.e. eating more) they quickly realize how much harder it is for them than continuing along with their restrictive behaviors. The following excerpt from a participant in the so-called "Anorexia Experience Study" discussed in Charland (2013) does a good job of describing this phenomenon on the basis of her own experiences:

> *For a long time I thought it was, there was nothing wrong with me, it was, there was nothing wrong with me, it was just other people thought there was, something wrong with them not me, but um . . . over the summer I did feel that I really wasn't in control of what I was doing and . . . it's sort of . . . before then I never really tried to get better, I'd always been forced to or, kind of, gone along with it to keep other people happy and I thought that as soon as I decided I did want to get better I'd be able to, but now I realize it doesn't quite work like that and so that's kind of made me see it as a bit more of an illness, something you don't have complete control over (p. 359).*

This is the sort of shift in experience we would expect, given my suggestion that anorexics experience illusory willfulness with respect to their food restriction due to the fact that their pathologically-driven behavior is egosyntonic and thus "matches up" with their considered intentions prior to pursuing recovery. Once food restriction is no longer fully egosyntonic, however (because the individuals have formed new intentions to eat more) the true nature of their food restriction is revealed to them. Given that the influence of the narrator module

in the integrated model is not meant to be insuperable, it is appropriate that with new evidence (i.e. the significant effort required to eat more) the contents of the anorexic's sense of agency would shift along with this new information gained.

In conclusion, the puzzle with which we began arises due to the fact that anorexic food restriction is egosyntonic. Because these actions are egosyntonic and the intentions and motives of the individual cohere with the behaviors that are being habitually selected, the anorexic falls victim to an illusion of willfulness that is made possible by the structure of the integrated model of the sense of agency. This proposal also provides a satisfying explanation as to why anorexics tend to realize they are no longer in ordinary control over their eating behavior only once they have begun to pursue recovery. It is my hope, then, that this proposal will help to shed light on the mechanisms underpinning anorexia nervosa that still remain poorly understood. Furthermore, I hope this account will serve as one example of the importance of carefully attending to the first-personal reports of those who experience the conditions we theorize about.

A HORSESHOE MODEL OF PATHOLOGICAL LOSS OF CONTROL

0. Introduction

Analyzing the so-called *loss of control* that occurs in addiction is difficult in part due to the fact that this phenomenon can appear either *familiar* or *alien* to us depending on what is emphasized. One method of inquiry that is commonplace in the philosophical literature (and especially in writing on action theory) is to begin by zeroing in on more familiar instances of failures of the will. The thought is that once this target has been adequately theorized about, we can gain a proper understanding of what goes wrong in addiction from the conceptual vantage point of universal experiences of akrasia, weakness of will, and the like. This method lends itself to conceiving of addiction in a way that is continuous with non-pathological failures of the will.

In contrast to this predominantly armchair approach, an alternative method of trying to understand loss of control as it pertains to addiction begins by examining the empirical data on what goes on in addiction from the perspective of psychopathology and the brain sciences. This latter method of inquiry tends to lead to an understanding of addictive behavior as fundamentally alien and separate from ordinary human agentive experience. Recently, however, even the former variety of theorizing about addiction has become increasingly up to date and intertwined with the empirical sciences. Given this development, it might seem difficult to determine which approach to the loss of control that characterizes addiction provides the best way forward: is the phenomenon more familiar to non-disordered human agentive experience, or is it more alien?

This paper offers a way forward that goes beyond the "familiar" versus "alien" dichotomy by casting a wider net with respect to the clinical data at our disposal. In particular, I highlight the empirical similarities and dissimilarities between substance use disorder (SUD) and anorexia nervosa (AN) in order to develop a model of pathological loss of control that respects key intuitions from both the "familiar" and "alien" camps in this debate.

In order to arrive at what I will call the "horseshoe model" of loss of control, a fair amount of philosophical and empirical ground must be covered. To that end, I will begin in §1 by providing an overview of two representatives of the "familiar" and "alien" approaches to understanding addiction. Then, in §2, I will delve into the empirical literature on anorexia nervosa as it pertains to the empirical literature on addiction. Finally, in §3, I will put forward a model that adequately represents the points of overlap and dissimilarity between the ostensible paradigm case of a lack of self-control (SUD) and of excessive self-control (AN). This model will share the intuition of the "familiar" approach that pathological loss of control is in some sense continuous with nonpathological behavior while also respecting the "alien" approach's intuition that certain key elements of pathological agentive behavior can only be understood in the context of empirical data.

1. Two Competing Accounts of the Relationship Between Addiction and Akrasia[12]

In this section I will lay out the points of similarity and disagreement between the recent accounts of addiction offered by Nick Heather (2016a, 2016b, 2020) and Edmund Henden (2013, 2016). I have chosen to focus on these two authors over others that address the question of whether addiction is a form of akrasia (e.g. Dill and Holton 2014, Butlin and Papineau 2016) because of the amount of overlap that is present in these two accounts that are otherwise at odds with one another. This overlap will allow me to highlight what I believe is truly at issue, namely the question of whether akratic action is more "familiar" or more "alien". Although I will go on to agree with Henden's account significantly more than Heather's, the model I will go on to develop in §3 will respect key intuitions from both authors. It is also worth mentioning before proceeding that both Heather and Henden are on board with the interdisciplinary literature that argues *against* the so-called "medical model" of addiction. According to the medical model, addicts *literally cannot do otherwise* when they succumb to temptation due to the presence of strong drug cravings that "hijack" their brains. In what follows, it will be assumed that this understanding of addiction actions is incorrect, and that literal compulsion in the philosophically-loaded sense is not present in addiction[15].

[12] I am aware of the fact that the two authors I have chosen to serve as foils to one another happen to have confusingly similar surnames. If you, like me, occasionally rely on mnemonic devices to keep things straight, might I suggest the following tricks to keep the authors' names and views clear: Heather comes before Henden alphabetically, and I introduce Heather's account before Henden's account in this paper. Also, the "a" in Heather stands for "akrasia", which is what Heather argues addiction is. [15] For further discussion, see, e.g., Pickard (2017).

1.1 HEATHER ON ADDICTION AS A FORM OF AKRASIA

To that end, let us begin with Heather's account of addiction, which tries to place the species of repeated actions that constitute addiction within the broader category of akratic actions. More precisely, Heather's view is that what he calls "ordinary akrasia" is a universally recognizable phenomenon that is not "qualitatively different" from the phenomenon of addiction. Although the form of akrasia that is operant in cases of addiction involves greater suffering and occurs with greater frequency and regularity than instances of ordinary akrasia, the addictive case is nonetheless meant to exist along a continuum of akratic actions that share fundamental features in common with one another. This core idea that we can understand addictive actions from the perspective of a universally recognizable human experience makes it easy for Heather to deny that there is a meaningful boundary between addictive and non-addictive actions. Although I will go on to disagree with Heather about the extent to which commonplace instances of akrasia can be instructive when analyzing addiction, I do view this lack of a clear boundary between normalcy and pathology to be a positive feature of his account that my own proposal will share.

This lack of a determinate border between pathological and non-pathological behavior goes hand in hand with Heather's tactic of emphasizing vignettes of agents exhibiting ordinary akrasia in order to make his case that the central features of these cases can also apply to addiction. The case that appears most centrally in Heather's (2016b) account is that of Mele's (1987) failed dieter, Fred. As the story goes, Fred is a man who has resolved to no longer eat an after-dinner snack for the month of January.

Up until the moment in question at which we are introduced to Fred, we are told that he has successfully resisted his regular after-dinner snack cravings by repeating to himself his reasons for forming a resolution to abstain from late night snacking in the first place. One fateful night, however, Fred is tempted with the visual cue of a slice of chocolate pie sitting in plain view as he opens his fridge for a beer. Faced with the desire to "throw in the towel", Fred internally rehearses his reasons in favor of abstaining from the slice of pie, but this time it is to no avail. Despite the fact that he has judged it would be best for him not to take and eat the pie, he proceeds to take it, slather it with whipped cream, and eat it. Poor Fred has fallen victim to what Setiya (2007) calls "clear-eyed akrasia" meaning that Fred acts knowingly against his own best judgment throughout the execution of his action.

In terms of Heather's application of this case, the salient detail is meant to be that Fred is a clear, *phenomenologically relatable* case of both diachronic and synchronic akrasia or weakness of will. As far as Heather is concerned, the relevant subsection of akratic action that can be extended to addiction is akrasia that is both diachronic and synchronic in nature. In other words, it conforms to Mele's (2012) and Davidson's (1980) synchronic criterion of going against an agent's all-things-considered judgment at the time of action, and it also conforms to Holton's (2009) diachronic criterion for weakness of will, which holds that the action in question must violate a previously formed resolution. Although Heather grants that there are likely to be instances that only satisfy either (synchronic) akrasia or (diachronic) weakness of will, he contends that there are at least

some key instances, such as the case of Fred, that satisfy both criteria[13]. And it is this category of akrasia, he claims, that is continuous with addiction.

What is addiction, then, according to Heather? After stipulating the synchronic and diachronic criteria he believes ought to be included within a satisfactory conceptual analysis of addiction, Heather (2016b) settles on defining addiction as "repeated and continuing failures to refrain from or radically reduce a specified behavior despite prior resolutions to do so" (p.141). Later, in Heather (2020), he adds the requirement of an interplay between short-term rewards and long-term punishments, which gives us "[a] repeated and continuing failure to refrain from or radically reduce a behavior that gives short-term rewards but longer-term punishments despite prior resolutions to do so" (p. 3). In support of this understanding of addiction, Heather cites clinical experience working with addictive disorders as well as his own experiences of struggling to quit smoking cigarettes.

Returning to the case of Fred, Heather's view is that "a paradigm case demonstrating the link between addiction and ordinary akrasia is the dieter who, when offered a slice of chocolate cake says, 'I know I shouldn't but I will' and then proceeds to take the cake and eat it" (p. 139). As far as I can tell, Heather's justification for this relies on the descriptive parallels he draws between Mele's "ordinary akratic" case of Fred and Heather's own preferred analysis of addiction. First, he calls attention to the fact that Mele's

[13] Although his target phenomenon could also rightly be called weakness of will according to Holton's terminology, Heather opts to only use "akrasia" from here on out so as to avoid adding further terminological confusion. I will also stick to using "akrasia" from here on out for the same reason.

analysis of Fred's volitional situation can also serve as an analysis of the situation many addicts find themselves in. In particular, the passage from Mele (1987) that Heather focuses on is one in which Mele is describing the case of Fred in order to make the point that intentional action against one's better judgment at the time of action can still be considered free. The emphasis on Fred's cognitive rehearsals of his reasons for not succumbing to the chocolate cake is meant to show that he *is*, in fact, very much aware that he ought not to take the cake and eat it. Likewise, the mention of his "carefully" spreading the whipped cream topping supports the claim that this is an intentional act carried out over a series of attentively performed steps.

In articulating what I take to be a central move in connecting Mele's Fred to that of addictive actions, Heather (2016b) writes,

> *Fred's behavior, I suggest, is not different in kind from what is normally considered addictive behavior and, if it were repeated, would certainly fit a part of my provisional definition above of a failure to refrain from (drug use) despite prior resolutions to do so… in terms of the nature of the akratic action itself it [Fred's behavior] is not qualitatively different from someone who has resolved to quit smoking on New Year's Day and has failed to keep that resolve, i.e. has relapsed to addictive behavior (p. 142, emphasis mine).*

In other words, the relevant actions of the relapsing cigarette addict and the failed dieting actions of Fred are of the same kind due to the fact that they share the essential features of consciously and intentionally violating a previously formed resolution that one still takes to be the best course of action. Although the strength and emotional significance of the addiction-related resolution, as well as the frequency of such violations, will be different

in the case of the addict versus Fred, Heather claims these two features are not significant in categorizing said actions.

At this point, it is important to call attention to the fact that the above considerations Heather articulates in favor of categorizing addictive actions as akratic actions take place only at the level of action-theoretic conceptual analysis and thus do not rely on or make use of any empirical considerations. However, Heather also cites a number of empirical findings from both addiction research and non-pathological psychology studies in order to orient his account within the broader interdisciplinary understanding of addiction and akrasia. It is important for Heather's account that his definition as well as his Mele-inspired analysis of addiction cohere at least to some extent to what the science has to say about addiction. Otherwise, it would be left vulnerable to the simple objection that the portrayal of addiction he is trying to connect to failed dieters does not in any way resemble the real-world phenomenon he is purporting to analyze.

What we get from his (2016b) chapter is, by his own admission, a "rough sketch of possible links" between his account and empirical research programs concerning addiction and other self-control related phenomena. These include behavioral economic theories of addiction related to hyperbolic discounting, dual process theories of action selection and control, Roy Baumeister's and his colleagues' work on ego depletion and cognitive control, and finally Holton and Berridge's (2013) article on addiction and weakness of will inspired by Robinson and Berridge's (1998) account of incentive salience and sensitization in addiction. Then, in his 2020 article, Heather delves further into how he

sees the concept of akrasia as fitting into a dual process theory of addiction. In §1.3, I will discuss each of these theories in relation to Heather's and Henden's respective interpretations of them. For now, however, I only intend to highlight the fact that these empirical connections are meant to complement and bolster Heather's understanding of addiction as akrasia, given that it otherwise relies heavily on armchair considerations and appeals to intuitive phenomenology.

1.2 HENDEN ON ADDICTION AS A MALFUNCTIONING OF THE WILL

Henden's (2016) chapter takes a similar starting point as Heather's in that he is after an understanding of addiction that can deny that addicts have literally lost free will over their behaviors when engaging in their addictions. His view departs from Heather's, however, in that he is interested in carving out a way in which he can do this while simultaneously denying that addiction is a form of akrasia or weakness of will[14]. Instead, Henden's view is that addiction ought to be categorized as a special kind of malfunctioning of the will that can be explicated via a dual process approach to decision making (to be discussed below). This is noteworthy given that Heather considers his own account of addiction as akrasia to be well suited for being framed alongside a dual process account.

What, then, does Henden disagree with Heather on when it comes to addiction? To start, Henden explicitly disagrees with Heather's claim that the frequency and regularity

[14] Although Henden generally speaks of "weakness of will" and Heather generally speaks of "akrasia", I take them to be referring to the same phenomenon that is at issue. I feel this is justified due to the fact that both authors take care to situate their views in the context of Mele's and Holton's understandings of akrasia and weakness of will, and both of them consider their accounts to be applicable to both Mele's and Holton's accounts.

with which addictive actions occur is theoretically unimportant. Contra Heather, Henden notes that it hardly makes sense to speak of "one off" addictive actions, which suggests that one cannot accurately account for the nature of addictive actions by zeroing in on a single addictive act. The idea is that, by widening our focus to the overall trajectory of addictive behaviors, we can attend to gradual changes that may ultimately alter the character of the relevant behaviors. Henden then suggests that this wider focus allows us to see that one thing that appears to be unique about addiction is that it is compulsive in the clinical rather than the philosophical (i.e. non-free will-trumping) sense. In his own words,

> *[Addictive] behavior is characterized as strongly cue-dependent in the sense that it is regularly triggered by certain situations, places or people associated with the type of behavior in question; there is a feeling of been driven again and again to behave in precisely that particular way (often in spite of oneself), and it is a common experience that resistance, however sincere, becomes increasingly difficult over time (Henden 2016, p. 11-12, emphasis mine).*

This emphasis on the diachronic and increasingly (clinically) compulsive nature of addiction can be wielded as an objection to Heather's style of analyzing addiction and ordinary akratic actions in isolation. In order to put flesh on this suggestion, however, Henden must appeal to much of the same scientific research on addiction that Heather cites. In the following subsection, I will go over how each author chooses to situate his preferred theory within the relevant empirical context while also noting the ways in which they diverge from one another.

1.3 HEATHER AND HENDEN ON THE SCIENCE OF ADDICTION

Hyperbolic Discounting:

Hyperbolic discounting, or the phenomenon of disproportionately discounting the utility of long-term rewards in favor of short-term rewards (Cf. Ainslie 2001), is a phenomenon that is by no means unique to addicted subjects but that appears to be particularly pronounced in addicted populations (Bickel et al. 2014). Heather takes this fact to be supportive of his theory, given that this tendency seems to map on quite nicely to the commonsense understanding of akratic action, and research shows that addicts are especially prone to it. However, Henden takes issue with reading too much into this fact. As he points out, the view that increased hyperbolic discounting is a constitutive feature of addiction relies on the assumption that failed attempts to abstain are always due to shifts in judgment on the part of the addict. While this is sometimes the case, he claims it is surely not always the case. Sometimes, addicts report acting on their cravings even while remaining fully cognizant of the fact that it would be best for them to abstain. At the very least, then, it would appear that hyperbolic discounting cannot be a necessary condition for addictive behavior.

"Wanting" vs. "Liking":

Both Heather and Henden reference Robinson's and Berridge's (1993) influential incentive salience theory of addiction. Very briefly, this theory claims that the dopaminergic (i.e. reward) system of an addict's brain becomes "sensitized" over time to drug-specific cues due to the direct manner in which addictive substances affect dopamine

production. Here, "sensitization" refers to the phenomenon by which high levels of dopamine are produced in response to exposure to drug-specific cues that eventually become disproportionate to and even dissociable from the subject's conscious evaluations of said drug. The result is that the reward system of the brain continues to "want" the substance in question (in a sense of "wanting" that is driven by dopamine levels rather than cognition) even when the subject no longer consciously "likes" the substance (Cf. Holton and Berridge 2013). This potential decoupling of "wanting" from "liking" is meant to explain why addicts crave (or "want") their drug of choice even when they no longer particularly "like" or enjoy it, as well as why it is so uniquely difficult for addicts to abstain.

Heather (2016b, 2020) does not say much about how exactly this theory of addiction can be interpreted in light of his own, although his discussion of dual process accounts (discussed below) does make use of attentional bias studies that are often connected to incentive salience-related cravings. For Henden's account, however, incentive sensitization is an important component that ultimately contributes to addictive actions being meaningfully different sorts of behaviors from non-addictive actions. This is because incentive sensitization is theorized as contributing to attentional bias, which is the tendency for one's attention to be directed toward a certain sort of (in this case, drug-related) stimulus (McKay and Efferson 2010). What this means is that addicts' brains perceive (either consciously or, oftentimes, unconsciously) drug associated cues across all sensory modalities as particularly salient and attention grabbing. Since directed attention is thought to work alongside executive control (more on this later) within the human decision-making system, this in turn makes it increasingly likely that addicts will

experience cue-driven cravings and will have their actions biased toward fulfilling their addictive cravings. And, because Robinson's and Berridge's theory predicts that long-term addiction will lead to an increasingly sensitized (i.e. larger) dopaminergic response, this accurately predicts the trend of it being generally harder to overcome an addiction the longer one has been engaged in it.

The Dual Process Approach to Addiction:

In order to fully appreciate how cue-driven attentional bias is implicated in addicts' decision-making capacities, we will now need to venture into the literature on dual process accounts of decision-making that have become popular within cognitive psychology, and in particular the dual process accounts that are focused on explicating addiction (e.g., Bickel and Li 2010, Wiers et al. 2016). As stated above, both Heather and Henden consider dual process approaches to be compatible with their respective accounts. The basic idea behind dual process theory is that the decision-making system is composed of two distinct types of processes: one is fast, associative, automatic, unconscious, and highly influenced by environmental cues and implicit learning (type-1 processes), and the other is slow, reflective, effortful, analytic, and responsive to higher order cognitions (type-2 processes)[15].

[15] As Heather (2020) points out, the basic structure that is argued for in dual process accounts of decision making is already anticipated in Davidson's (1982) work on the possibility of akrasia. Davidson's claim that semi-autonomous mental structures are needed in order to explain how an agent can knowingly act against her all-things-considered judgment is incredibly reminiscent of the basic tenets of modern dual process theory.

In healthy individuals, the two types of processes work together in order to arrive at rational decision making. This happens when type-1 processes are able to select appropriate and useful information for the context the agent happens to find herself in, while type-2 processes are able to both (i.) successfully inhibit any impulsive tendencies of the type-1 process responses and (ii.) use the input from type-1 processes in order to form judgments and implement actions that are in line with the agent's goals. According to Heather's (2020) preferred interpretation of dual process theory as applied to addiction, addiction is the result of the dynamic interactions between type-1 and type-2 processes becoming dysregulated. In his own words, "addiction involves a failure of top-down regulatory control of bottom-up automatic processing, a failure due to an unusually strong impulsive system, an unusually weak reflective system, or a combination of both" (p. 6). Strikingly, this gloss comes quite close to what Henden will go on to say about his own interpretation of dual process theory as it pertains to decision making in addiction. It is important to note, however, that even in his terminology Heather's description of the dysregulation that takes place is relatively noncommittal as to its cause—it speaks only of a "failure" of inhibiting an unusually strong impulsive system and of a potentially "weak" reflective system. It says nothing about how this state of failure or weakness comes to be. The upshot for Heather is that this same analysis of a dysregulated decision-making system could presumably be applied to non-pathological failed dieters such as Fred.

Henden, on the other hand, makes use of this same body of research in order to highlight his main point that the breakdown between type-1 and type-2 processes that occurs in addiction is substantively different from whatever is happening in garden variety

akratic agents like Fred. Henden focuses on two elements in motivating this claim: i) the abnormal strength of attentional bias toward drug-related cues in addicts, and ii) the psychological research on the so-called "limited resource model" of cognitive control. Regarding the former, he distinguishes between two ways in which drug-related attentional biases might impede the functioning of type-1 and type-2 processes. First, the excessive attention may affect type-2 processing such that the agent overappreciates considerations in favor of drug-related actions. This could lead to some form of hyperbolic discounting of other long-term goals, although in cases of so-called "willing addicts" this need not be the case. I will reserve further discussion of the phenomenon of willing addicts for §3, but for now I will note that Henden supposes this to be a less prevalent occurrence in addiction than the more obvious effects of attentional bias on type-1 processes.

According to Henden, drug-related attentional biases on type-1 processes create an environment in which the addicted subject is particularly vulnerable to what is referred to by Baumeister and colleagues (Cf. Muraven and Baumeister 2000) as ego depletion. Although there have been issues in recent years regarding the reproducibility of Baumeister's proposed physiological mechanism through which ego depletion occurs (Cf. Kurzban 2010), the research program's central claim that tasks which require sustained, directed attention and focus (typically referred to as "executive control" or "cognitive control") appear to draw upon a limited resource is not affected by these concerns. The phenomenon of ego depletion itself, which has been replicated over one hundred times (Inzlicht and Schmeichel 2012), occurs when subjects who have already completed tasks requiring cognitive control perform worse than control subjects on subsequent cognitive

control tasks. This is relevant to the case of addiction because drug-related attentional bias due to incentive sensitization is exactly the sort of thing one would expect to deplete one's finite resources of cognitive control within a given period of time. And, indeed, Henden cites drug-related Stroop task studies such as Cox et al. (2006) that have been interpreted as evidence of cognitive depletion in addicts versus controls due to drug-related attentional bias[16].

Putting it all together, drug-related attentional bias caused by incentive sensitization, and the effect this can have on addicts' capacity to exert cognitive control in the context of a dual process theory, are meant to set addicts apart from akratic agents. Whereas Heather focuses on the ways in which the dual process theory might lead to incontinent or irrational action in ordinary akratic cases and then extends this to the case of addiction, Henden considers the unique psychological and neurological situation that addicts find themselves in vis-à-vis incentive sensitization (and its downstream effects) to be sufficient cause for categorizing addiction as something unlike the familiar case of akrasia. Instead, he opts to classify it as a "malfunctioning of the will", which he takes to be a more accurate description of the phenomenon at issue.

[16] Stroop tasks are used in experimental settings to measure the Stroop effect, which is the amount of time it takes to successfully complete a task that involves inconsistent stimuli. In this particular instance, patients with SUD and healthy controls were asked to name the various colors of drug-related and drug-unrelated words. Studies such as Cox et al. (2006) have found that there is greater temporal delay for addicts with respect to the drug-related words but not for healthy controls. The standard interpretation of increased temporal delay or "Stroop interference" is that it is a result of increased attentional bias, since completing the activity quickly and accurately requires sustained directed attention.

At this point, one might object that "malfunctioning of the will" is a nonspecific and noncommittal term that is not obviously mutually exclusive with Heather's concept of akrasia, and in response I must admit that I am more or less inclined to agree if we were to consider each of these accounts in isolation. In the following sections, however, I will show that Henden's "malfunctioning" account is more easily adaptable in order to account for the heterogeneity of actions that appear to fit the empirical profile of the phenomenon both authors are interested in. Furthermore, it is important to note that Heather and Henden come away from their analyses of the dual process theory as applied to addiction with different takeaways. Heather sees the mechanisms that may lead to incontinent action in addiction as conforming to general features of human akratic action, whereas Henden sees the neurological and psychological profile of the addict in such situations as constituting a conceptually distinct sort of phenomenon.

Finally, Heather repeatedly links his own account of addiction to the familiar phenomenology of akratic action as exemplified by agents such as Fred. Henden, however, does not rely on any particular phenomenology or any such appeal to a familiar experience. This is presumably because Henden explicitly mentions and acknowledges the existence of willing addicts (Cf. Flanagan 2013). Since willing addicts need not share anything phenomenologically in common with Fred as they engage in drug-oriented behavior, any appeal to recognizable phenomenology would arguably be superficial and potentially misleading (more on this later). In what follows, I will show that this non-committal stance with respect to the associated phenomenology of actions involving a lack of self-control is

another feature of Henden's account that is vindicated by further consideration of the relevant psychopathology literature.

2. Empirical overlap between disorders of deficient and excessive cognitive control

The empirical factors that both Heather and Henden consider in formulating and defending their respective analyses of addiction fall under two general categories. On one hand, we have factors (i.e. incentive sensitization and its application to the dual process approach) that are connected in one way or another to attentional biases toward drug-related stimuli that are disproportionately weighted for addicts at the neurological level. On the other hand, we have findings that seem to suggest a general tendency toward impulsivity outside of the direct domain of their chemical addictions, as evidenced by the increased hyperbolic discounting observed in addicted populations. In this section, I will outline the ways in which the former category of empirical considerations overlaps to a surprising extent with the research findings of patients with anorexia nervosa (AN). In addition, the latter category of findings that seems to indicate a general tendency towards impulsivity in addicted populations will be contextualized within the overall structure of similarities and differences between disorders associated with heightened impulsivity and those marked by excessive rigidity.

Anorexia nervosa is a debilitating mental disorder that is characterized by, among other things, excessive dietary restriction as well as other compulsive behaviors aimed at achieving weight loss and maintaining a low body weight, which is a goal that anorexics consider to be highly valuable (see Paper 1 and Paper 3 for further discussion). It is also a

disorder that occupies a unique position within the philosophical and psychological discourse regarding self-control and addiction in that it involves a symptomology that is *prima facie* opposite to that of SUD. However, a closer examination of our current understanding of the psychological and neurological factors at play in AN reveals that the disorder shares more in common with SUD than a commonsense understanding of the condition might assume. That being said, anorexics do display features that are the opposite of those exhibited amongst addicted populations by and large, such as their performance compared to healthy controls in studies measuring hyperbolic discounting. After outlining these similarities and differences in the present section, the aim of §3 will be to propose a model of self-control that accurately reflects the complex relationship between these two disorders.

In order to understand the connections between addiction and anorexia we must first consider in more detail the psychological concept of cognitive control. Cognitive control, which Brooks et al. (2017) note is variously referred to across the literature as "cognitive inhibition, affect regulation, self-regulation, top-down control, and cognitive-emotion interaction" (p. 1), is often used as an umbrella term for top-down processes that regulate and control the selection and initiation of goal-directed actions (Henden 2016). In terms of its relation to the dual process model of human decision making, cognitive control is a primary capacity of the executive type-2 processes that can be realized to a greater or lesser extent in order to control the more impulsive tendencies of type-1 processes. It is also the faculty that is considered to be a depletable, finite resource according to the ego depletion literature.

At first glance, the literature measuring the relative performances of anorexic and addicted populations on tasks designed to test cognitive control is exactly as one might expect. That is, anorexics exhibit above average cognitive control compared to healthy controls, whereas individuals with substance use disorder perform worse than healthy controls (Bickel and Marsch 2001, Bickel et al. 2014). For example, Steinglass et al. (2012) used an intertemporal choice task to measure the relative rates of hyperbolic discounting among anorexic patients and healthy controls. The task was designed to measure the rate at which individuals begin to discount the value of greater monetary rewards to be gained sometime in the future over smaller monetary rewards to be received immediately. The tasks had real-life consequences (i.e., participants actually received Amazon gift cards for the stipulated amount at the stipulated time) so as to promote sincere and realistic responses. Steinglass and colleagues found that anorexic patients valued far-off monetary rewards significantly more than healthy controls, which is to say that they exhibited a *decreased* rate of hyperbolic discounting as opposed to th*e increased* hyperbolic discounting rates commonly exhibited by addicted populations when performing such tasks. These results are noteworthy in that they indicate a general, non-disorder-specific tendency toward increased gratification delaying capacities among AN patients and increased impulsivity among SUD patients.

Experimental results such as these are often taken to reflect above-average cognitive control capacities in anorexics in contrast to below-average cognitive control capacities in addicts. And, to a certain extent, this does appear to be the case. Digging deeper, however, things quickly become more complicated once we begin to explore the

mechanisms by which anorexics achieve this above-average performance on certain measures of cognitive control. The first thing to note is that anorexics do *not* perform better than healthy controls on *all* of the distinguishable facets that contribute to overall executive functioning. In particular, they perform below-average on "set-shifting" tasks that are designed to measure an individual's ability to flexibly "shift" between different tasks according to situational demands (Cf. Steinglass et al. 2006, Tchanturia et al. 2004). This is commonly thought to be due to the excessively rigid and over-controlled decision-making style observed amongst anorexics, which leads to an over-reliance on entrenched behavioral patterns over novel and adaptable responses to changing stimuli. This pattern even appears to extend past ostensible recovery, given that King et al. (2019) found that individuals who had recovered from AN exhibited higher accuracy but slower response speed during set-shifting tasks when compared to healthy controls.

While interesting in its own right, the fact that anorexics do not perform better than healthy controls in *all* measures relating to executive functioning does not on its own affect the initial assumption that AN and SUD are diametrically opposed disorders in this regard. However, the means by which anorexics appear to achieve high levels of cognitive control is also implicated in their poor performance in set-shifting tasks, and it is here that we can begin to see commonalities between the two disorders. The points of intersection and dissimilarity between AN and SUD have recently been the focus of work by Samantha Brooks and colleagues (Brooks 2016, Brooks et al. 2012, 2017), and some of the insights outlined in these articles will serve as building blocks for the model

I will go on to propose in the following section. Brooks et al. (2017) argue that increased "epistemic foraging" (i.e., cognitive sampling of external and internal cues that are relevant to one's decision) is involved in the elevated cognitive control capacities of anorexic subjects. In particular, Brooks et al. suggest that anorexics make use of the very cognitive ruminations pertaining to eating and weight loss that are characteristic of anorexics, in addition to any disorder-friendly cues in the environment, as a means to keep attention focused on disorder-congruent stimuli over stimuli that are disorder-incongruent (e.g. calorically dense food). In fact, Brooks et al. suggest that further study of this effective distraction mechanism as it takes place in AN may prove useful for developing treatments for SUD as well as binge eating disorder, since telling addicts to "try to focus on something else" when presented with drug-oriented stimuli is a beneficial recommendation that Heather (2020b) discusses in some detail. Harkening back to the case of Fred, recall that it was his failure to maintain his strategy of rehearsing reasons not to eat the pie that led to his akratic lapse. In Brooks et al.'s terms, Fred's epistemic foraging abilities were not sufficient in that instance.

Brooks et al.'s suggestion that the mechanisms underpinning one disorder might be relevant to the treatment of the other speaks to the similar channels through which anorexic and drug addicted behavior is maintained. Anorexics appear to refrain from their physiological urges to eat by successfully fixating on their own disorder-relevant stimuli, whereas addicts fail to abstain from drug use in part due to their sensitized attention toward drug-associated stimuli. Of course, anorexic abstention from nourishment is not a beneficial use of cognitive control faculties but rather a maladaptive one. It is along these

lines that Brooks et al. (2012) advocate for a "spectrum model" of eating disorders in which the excessive cognitive control of AN occupies one extreme end of the spectrum and the extreme impulsivity (reminiscent of SUD) of binge eating disorder occupies the other end, with bulimia nervosa in between the two[17].

The relevant overlap between AN and SUD goes deeper, however, than the mechanism of one disorder potentially being useful in the development of an antidote for the other. In order to see why this is the case, we must attend to the function of what I have been referring to as "disorder-relevant" or "disorder-congruent" stimuli in anorexia nervosa. As has already been discussed, these stimuli can include external sensory stimuli (e.g., photos of emaciated bodies posted in online forums as "thinspiration" or "thinspo" that many anorexics seek out to "keep themselves on track"), as well as cognitive ruminations relating to weight loss (e.g., "I will only consume food *x* up until the caloric amount *y* today). Thanks to increased epistemic foraging and their ability to sustain directed attention at an elevated rate, it is thought that anorexics are able to distract themselves from the innate physiological cues that are pulling them in the opposite direction of action (i.e. to eat more food when hungry). As Brooks et al. (2017) describe it, these cognitive ruminations "may engender the episodic representation of images evoked by deliberative prefrontal cortex predictive processes, such that internally generated images

[17] More precisely, the pure-restriction subtype of anorexia nervosa would occupy one extreme end of the spectrum, with the binge-purge subtype of anorexia nervosa occupying the position in between pure restriction AN and bulimia nervosa. Although I have been referring to just "anorexia nervosa" throughout, this should be taken to be referring to the pure restriction subtype of AN. In the model I will go on to propose, then, pure-restriction AN will similarly be on the more extreme end compared to the binge-purge AN subtype.

are eventually furnished with a saliency akin to a concrete object" (p. 11). Although the supposed saliency and causal effect of these ruminations may seem extreme, it is worth remembering that we ought to expect an extreme first personal experience associated with a condition that involves pathologically elevated cognitive control and directed attention toward such cues.

There is more to the story, however, when it comes to the functions of disorder relevant cues in AN. This is because disorder-relevant cues as well as disorder-congruent behaviors are believed to be *intrinsically rewarding* to anorexic individuals (Keating 2012). In this way, disorder-congruent behaviors in AN parallel disorder-congruent behaviors in SUD, although the behaviors in question for AN are not ones that are transparently rewarding from the outsider's perspective. As O'Hara et al. (2015) note, it was once thought that anorexic food restriction was maintained by a general state of anhedonia wherein anorexics are unable to fully experience the rewarding psychological states associated with palatable food consumption. Thanks to significant developments in anorexia research over the past several years, however, it is now believed that anorexics can, in fact, experience high levels of reward but that their experience of what is rewarding is "contaminated". The so-called "contamination theory" as it pertains to AN posits that anorexics experience otherwise rewarding stimuli (such as nutrient-dense and palatable food) as threatening and aversive, and that otherwise unpleasant stimuli (such as depictions of emaciated bodies and the feeling of hunger) are experienced as desirable and rewarding (Keating 2010, 2012). In fact, Brooks (2016) indicates that the "impulse control spectrum" model of eating disorders as put forward by Brooks et al. (2012) fits well with the literature

which suggests that the exercise of cognitive control is itself perceived as intensely rewarding in anorexia nervosa, thereby leading to it becoming "an addiction in itself" (p. 8).

Brooks (2016) notes that such a set-up would lead to a *lack* of cognitive control for anorexics over the disorder-relevant compulsions themselves, a prediction that gels with the notoriously poor remission rates of individuals attempting to recover from AN. Furthermore, Keating (2010) argues that patients are unlikely to recognize that they have undergone reward contamination, which may in turn make it difficult for individuals to regulate and control their own behaviors once these processes are underway[18]. In fact, many of these considerations come very close to the justifications Henden (2016) uses in order to argue that abstaining from addiction-related behavior is importantly different from standard akratic behavior on the basis of research findings related to cue-driven attentional bias.

At this point, although a proper treatment of the intricacies surrounding anorexic reward processing is beyond the present scope (but see also Paper 1), the foregoing discussion has hopefully highlighted the salient points at which addiction and anorexia nervosa intersect. Below, I have included a table that highlights the points of similarity and

[18] In particular, Keating et al. (2012) write, "Patients, however, may not recognize that they are contaminating aspects of reward with punishment, due to overlapping neurocircuits that process reward and punishment (e.g. dopamine), which may facilitate neural and behavioral reinforcement, thus impairing patients' ability to regulate their behaviors" (p. 568). In other words, Keating et al. appear to be highlighting that a lack of awareness into the true psychological processes underpinning their food restriction may make it hard or even impossible for anorexics to cease their behavior, at least until greater awareness is gained. For a theory of impaired awareness in AN that lines up quite nicely with this prediction, see Paper 1.

dissimilarity between SUD and AN in reference to the experimental and clinical measures discussed in §1 and 2.

Empirical features	SUD compared to AN
Cognitive control measures related to hyperbolic discounting	**Dissimilar**—AN is above-average and SUD is below-average compared to healthy controls
Cognitive control measures related to set-shifting	**Similar**—Both disorders exhibit deficiencies relative to healthy controls
Disorder-relevant attentional bias having a significant effect on decision-making	**Similar**—Independent evidence on this for both disorders
Cognitive control intrinsically rewarding?	**Dissimilar**—Yes for AN, no for SUD

Table 1: A comparison of the empirical features associated with substance use disorder and anorexia nervosa.

3. The Horseshoe Model of Loss of Control

This paper began with a discussion of the points of overlap and disagreement between Heather's account of addiction as a form of akrasia and Henden's account of addiction as a (non-akratic) malfunctioning of the will. The essence of disagreement between these two accounts is, at the end of the day, a difference of emphasis. Heather argues elsewhere (2016a) that it is the behavioral level, rather than the neurological or psychological levels, that is essential to the phenomenon of addiction. A proper treatment of this argument is beyond the scope of the present paper, but it is worth noting how this emphasis on behavior lends itself to defining addiction solely in terms of an action-theoretic concept such as akrasia. Henden, for his part, chooses to emphasize the

phenomenon of loss of control as a means of understanding addiction. This differs from Heather in that "loss of control" can be arrived at via numerous causal pathways, which in turns lends itself to analyzing the neurological, psychological, and behavioral factors all together, as opposed to favoring one layer of explanation as Heather does.

One result of Heather's privileging of "repeated, failed attempts" as an essential feature of addiction is that he is forced to make somewhat awkward concessions in light of counterfactual considerations. Recall that his full definition of addiction that he arrives at in Heather (2016b) is "repeated and continuing failures to refrain from or radically reduce a specified behavior despite prior resolutions to do so" (p.141). At face value, this definition seems immediately vulnerable to an objection citing the existence of willing addicts, which is something Heather (2016b, 2020) does not engage with but that is addressed in Heather (2016a). Here, Heather responds to an objection Ole-Jørgen Skog (2003) raised in response to Heather's (1998) article in which he first defended the definition of addiction that is at issue. Skog rightly points out that it seems wrong to deny that individuals who we might otherwise want to classify as addicted are only addicted *if* and *when* they try to abstain from their behaviors and fail to do so.

In making his point, Skog uses an example of a longtime drug user who has no desire or inclination to quit using drugs at t_0 but then changes his mind at t_1 due to changing circumstances. At this point, it is reasonable to suppose that such an individual would find it difficult to quit and would likely experience some failures in trying to abstain. According to Heather's own preferred definition of addiction, we would have to say such a person was *not* an addict at t_0 but then became an addict once he tried to quit at t_1. Skog (justifiably,

in my view) finds this highly counterintuitive, and on this basis he suggests that Heather's definition fails to characterize the fundamental features of addiction. It is worth noting here that Heather (2016a) does in fact concede in response to Skog's objection, albeit "reluctantly", to use his own phrase. Heather opts to retain his preferred definition going forward without any counterfactual qualification but, in his own words, "with the understanding that [he] would concede if the counterfactual objection were made" (p. 13). My own view is that Heather's admitted reluctance to fully take on Skog's counterfactual objection is telling. In particular, I would like to suggest that this worry is cuing in to the fact that Heather's attempt to analyze addiction in terms of a singular occurrent akratic action is missing the mark in terms of what is truly essential to the phenomenon.

Henden's (2013, 2016), account of addiction, on the other hand, faces no such issue with Skog's objection. In fact, he is fully aware of Skog's (2003) counterfactual analysis of addiction and fully takes it on board as compatible with his own view. In a footnote, Henden argues that the important difference between an addicted and a nonaddicted drug user "resides in certain counterfactuals being true when the person is an addicted user and false when he is a non-addicted user" (p.11). To Henden, the relevant counterfactuals concern how each person would behave if his supply of drugs were to wane as well as how each user would react if his drug use were to begin causing tangible harm to other areas of his life. He says, "were his drug use to become associated with displeasure, emotional distress, or health problems, it would be true of the addict but false of the non-addict that he would continue to consume the drug, often experiencing a physical compulsion to do so" (Ibid).

It is at this point that we can clearly see how Henden's choice to emphasize the more general and flexible phenomenon of loss of control allows him to evade objections that present as more challenging to Heather's account. In my own view, the phenomenon of "loss of control" in its most relevant form is already a deeply counterfactual concept. As far as pathologies of agency are concerned, what is relevant is not whether an individual happens to lose control in a given circumstance. Rather, what is truly theoretically pertinent is whether there exists a longstanding pattern of psychological and neurological processes such that the agent is systematically vulnerable to a loss of control, regardless of whether this reality is in fact revealed to her at the present time. It is with this understanding of the concept of loss of control that I invite the reader to revisit Henden's (2013) description quoted above on how to counterfactually distinguish the addict from the non-addict. Again, Henden writes, "were his drug use to become associated with displeasure, emotional distress, or health problems, it would be true of the addict but false of the non-addict that he would continue to consume the drug, often experiencing a physical compulsion to do so" (p.11). Based on the empirical understanding of anorexia nervosa discussed in §2, I would like to suggest that an analogous counterfactual can be used to distinguish an anorexic individual from a non-pathological dieter. In order to properly characterize this point, however, it is time to finally introduce the "horseshoe model" that I will argue accurately captures the philosophical and empirical data on loss of control discussed thus far.

This model of self-control and loss of control that I am putting forward makes use of a metaphor from the so-called "horseshoe theory" that originated in political philosophy

and political science. The basic idea behind this metaphor is that a model of something that is commonly thought to be well-represented by a spectrum is in fact better represented by a "horseshoe" shape in which each "end" is closer to the other end than either end is to the center. By applying this spatial metaphor, one can argue that the two polar "extremes" of a certain concept or entity are in fact more alike than the points that occupy the "middle"[19]. I believe the most accurate and illuminating model of self-control (and lack thereof) in human behavior is not a spectrum (as Brooks et al. 2012 might suggest) but rather a horseshoe. In contrast to the sort of spectrum model that Brooks et al. (2012) might propose if they were to develop a model not just for eating disorders but also for SUD, the model I am envisioning would represent the impulsive end (where SUD and BED would be placed) as "closer" to the pathologically over-controlled end (where AN would be placed) than to the central "normal" region of the horseshoe shape.

The central motivation for adopting what I will call the horseshoe model of loss of control is that the closeness of the two extreme ends (call these the "impulsive" and "compulsive" ends) in the spatial metaphor can serve as a stand-in for the theoretical "closeness" of each extreme that stems from their shared counterfactual properties. I concur with Henden that a key distinguishing characteristic between a casual drug user and a drug user who can aptly be labeled as "addicted" resides in whether they would be able to stop with relatively minimal effort if they were to form an intention to do so. I therefore disagree with Heather's claim that repeated failures to abstain are a necessary and theoretically deep

[19] It should be noted, however, that I do not personally agree with the claims that are typically made with reference to the original horseshoe theory as it pertains to the spectrum of political ideologies.

feature of the lack of control that is operant in addiction. Rather, the theoretically deep feature is that a lack of control *would* reveal itself if the proper set of circumstances were to be realized. In this way, the two "ends" of the horseshoe have a property that the middle region lacks, *viz.* a lack of control at the counterfactual level.

One may well wonder at this point why I am choosing to privilege a counterfactual property that is more metaphysically complicated than Heather's "repeated failures to abstain". By embracing this (admittedly complex) property, however, we can get at the "nub" of what connects willing addicts, unwilling addicts, and anorexics who are "addicted", so to speak, to exercising cognitive control in the service of disorder-congruent behaviors. In terms of the willing addict, I believe Skog (and Henden) are correct in drawing attention to the fact that such an individual exhibits a certain form of loss of control that is grounded in the fact that her current lack of conflict is *not* due to the potential flexibility of her actions, but rather due to the fact that her current pathological behavior happens to be in line with what she reflectively wants to do at that time. The same can be said for anorexic individuals, who only realize their lack of control over their dietary behaviors once they begin to attempt recovery (Cf. Paper 1).

In support of this connection between the impulsive and compulsive ends of the Horseshoe Model, I find this line from Henden (2016) to be particularly illuminating: "Compulsivity and obsession—despite their superficial appearance of 'too much control'—seem on a deeper level to indicate the opposite of control" (p. 127). He goes on to say that "even if some addicts have stable preferences, all their beliefs and desires will

still be infused by drug-associated attentional bias; hence, by taking these beliefs and desires as inputs, their *practical reasoning* itself will in a sense be 'out of control' (Ibid.). Although Henden is referring here to willing addicts, the same observation can be very suitably applied to anorexics. After all, anorexic behaviors are theoretically striking precisely *because* the pathological actions are congruent with and even emblematic of the agent's desires and longstanding beliefs. It is only when one's desires and beliefs change from what they have been that one's lack of control over the relevant behaviors become apparent, which is the basis for the counterfactual similarity between the extreme compulsive and impulsive ends of the Horseshoe Model.

With all of these considerations in mind, I will now present horseshoe model in more detail. In what follows, please consult *Figure 1* (below) for a rough illustration of what is being proposed.

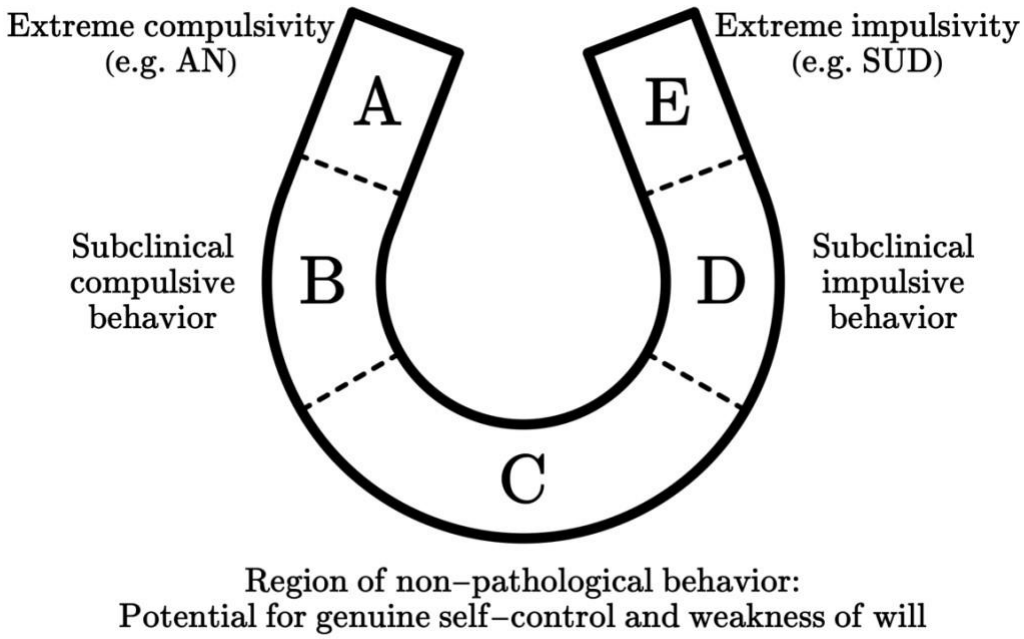

Figure 1: The Horseshoe Model of loss of control across pathological and non-pathological human behavior

As I envision it, at one end of the horseshoe (Region E) would be those disorders that are standardly categorized as the "high impulsivity" disorders: substance use disorder, binge eating disorder, and other behavioral addictions (e.g. gambling addictions) that are clinically associated with impulsivity. On the other extreme end of the horseshoe (Region A) would be those disorders that are commonly associated with excessive, pathological self-control, which is typically characterized in the literature as high compulsivity. Anorexia nervosa is the most obvious disorder to place on this end, and I believe obsessive compulsive personality disorder (OCPD), which happens to be highly comorbid with AN, is another plausible candidate.

Continuing on with the model, the "middle" region of the horseshoe (Region C) is the bread and butter of most action theorists, and it is the region that has generally

monopolized contemporary philosophical inquiry relating to willpower and weakness of will (e.g. Davidson, O'Shaughnessy, Mele, Holton, and also Heather). Region C is unique in that it is *potentially* immune to the counterfactual objection discussed above[20]. Finally, to the left and right of Region C lies Regions B and D, respectively. These two regions correspond to the human behaviors for which some of the clinical considerations relating to attentional biases etc. are present, but not to the extent that these factors are present in Regions A and E. In other words, the extent to which these individuals would struggle to cease their relevant behaviors would be greater than one would find in Region C but lesser than in Regions A and E.

An example of a type of behavior occupying Region B would be an individual who is incredibly strict with her diet and "clean" eating and exercise regimen. It may be that her neurological and psychological profile is such that she would not find it nearly as difficult as someone properly fitting the AN diagnosis to interrupt her dietary and exercise regimen, but it would nonetheless be rather difficult for her. Crucially, most individuals in Region B will generally be viewed by themselves and others as having strong willpower and above-average self-discipline. I personally think this type of individual is quite common in certain social circles and professions, including academia. It is also a region that I suspect many anorexics might "start" in before "moving" toward the extreme end of Region A, all the while believing that their behavior still corresponds to what I am calling Region B.

[20] I should note, however, that my intuition is that Region C is far smaller than what is depicted in Figure 1, and that most (if not all) of ostensibly "normal" human behavior ought to be subsumed by Regions B and D. Since I will not be defending this here, however, I have included Region C so as not to distract from what is currently being argued for.

In Region D, then, one could place individuals who are mild to moderate "problem drinkers" but are not at the point at which intervention or treatment is needed. Many individuals may spend some time in Region D during their early adult years before "aging out" and moving closer toward Region C, although some will instead progress further toward Region E[21]. Another way of describing regions B and D is that they represent behaviors relating to compulsivity or impulsivity that are "subclinical", which is to say that these behaviors would become clinical (i.e., diagnosable according to the Diagnostic and Statistical Manual of Mental Disorders) if they were to become more frequent or pronounced.

It must be stressed, however, that I do *not* view the dotted lines in Figure 1 as comprising real ontological boundaries between the regions just described. Indeed, a more accurate (though less informative) diagram for the horseshoe model would be one in which there are no determinate regions and there are labels designating the variation of counterfactual self-control capacity along the vertical axis. For this alternative representation of the horseshoe model, see Figure 2. That being said, I do think it is worthwhile to "section off" regions for the purposes of elaborating on the criteria for inclusion in different areas on the horseshoe.

[21] It is worth highlighting that the relations between Regions B and D and their respective extremes (i.e. Regions A and E) appear to be asymmetrical as a matter of how people are "out there in the world". That is, it seems to be the case that agents in Region D are more likely to move into Region E at some point over the course of their lives than agents in Region B are likely to move into Region A. The fact that substance use disorder and binge eating disorder are much more common than AN, and that binge-eating disorder is by far the most common eating disorder in general, supports this observation. An avenue of further inquiry would be to delve into exactly *why* this asymmetry seems to exist in the human population.

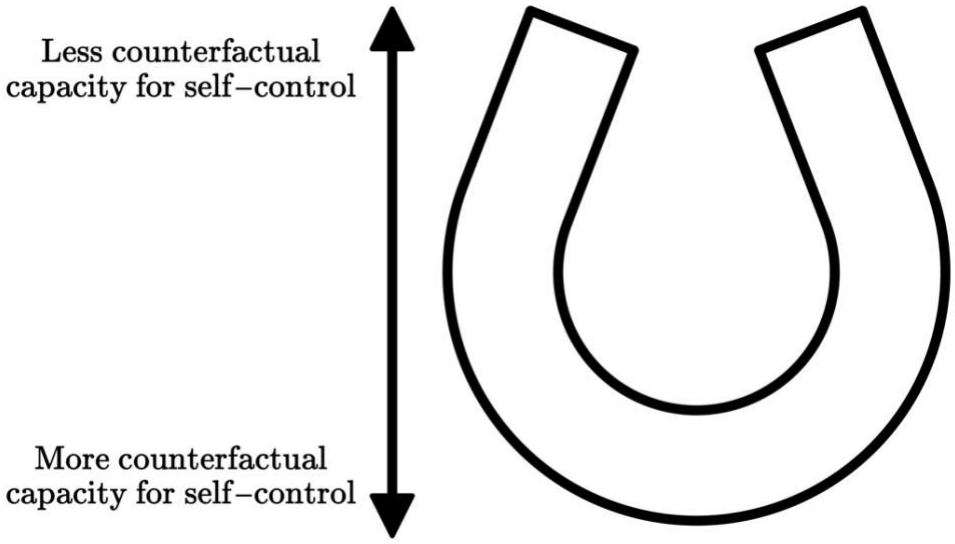

Figure 2: A simplified form of the Horseshoe Model demonstrating the variable of (counterfactual) self-control capacity as it varies along the vertical axis.

Despite the deep and meaningful similarities between the impulsive and compulsive extremes of the horseshoe model, I do not mean to suggest that they are effectively the same. For one thing, the compulsive end of the spectrum is unique in that it is marked by the "superficial appearance of 'too much control'", to use Henden's phrase. I believe it is this superficial appearance that has led so many philosophers (including Heather) to focus on the impulsive end of the horseshoe at the expense of the compulsive end. This is perhaps unsurprising given the fact that much of the literature on loss of control, akrasia, and willpower tends to rely on phenomenological analyses such as Mele's failed dieter Fred.

Recall that Fred's vignette relies on the familiar phenomenology of failing to stick to a resolution in order to appear plausible, and that Heather utilizes this commonsense

analysis and extends it to the case of addiction. It is for this reason that in §1.1 I made sure to highlight Heather's (2016b) central claim that the case of Fred is not "*qualitatively different*" from that of an addict. Although he does not explicitly spell out what he means by qualitative difference, Heather's points of emphasis when describing the case clearly focus on the *agentive experience of akrasia* as exemplified by Fred. Given that ordinary akrasia is Heather's theoretical home base and his vantage point through which he views addiction, we are now in a position to see the ways in which this perspective is limiting.

Indeed, any such account that centers its analysis of pathological loss of control around ordinary akrasia and its associated phenomenology is likely to leave the compulsive end of the horseshoe model out entirely. This is because the compulsive end of the horseshoe (i.e. Regions A and B in Figure 1) is not characterized by or directly associated with the *subjective feeling* of loss of control. In fact, it is often quite the opposite that is the case. An interesting study by Birgegard et al. (2009), for example, measured initial self-image variables as measured by the Structural Analysis of Social Behavior Model (SASB) against treatment outcome variables according to a 36-month follow up. In the AN group, the researchers found that the SASB self-image variable of self-control was the variable with the second-most predictive power of prognosis at 36month follow-up, with the first most predictive variable being the "baseline" measure of eating disorder severity (i.e., how sick the patient was) at the beginning of the study.

This led Birgegard et al. to conclude that "self-control is central to AN pathology" (p. 527), which is in keeping with Fairburn et al.'s (1998) theory of anorexia nervosa as being

centrally dependent on both the self-image and self-control. In other words, it appears that the agentive experience of high self-control, as opposed to a loss of control, is associated with the left-most region of the horseshoe model.

Finally, the horseshoe model of loss of control accurately predicts Pinto et al.'s (2014) findings that the capacity to delay reward in the context of temporal discounting measures robustly differentiated (egosyntonic) obsessive compulsive personality disorder (OCPD) from (egodystonic) obsessive compulsive disorder (OCD), the former of which is associated with a personality type and decision-making style emblematic of extreme perfectionism and rigidity. Pinto and colleagues found that although psychosocial quality of life measures were significantly impaired in both OCPD and OCD participants, OCPD patients exhibited significantly *less* temporal discounting (i.e. were more able to delay reward) than OCD patients and healthy controls, whereas OCD participants exhibited slightly more impulsivity than healthy controls. Given that OCPD is innately tied to heightened perfectionism and rigidity, the authors of this study noted that their findings are remarkably consistent with the literature that connects AN to both reduced temporal discounting as well as excessively perfectionistic and rigid behavioral styles. Given that OCPD and AN are highly comorbid disorders that appear to be significantly intertwined in their relation to excessive self-control, my claim that the leftmost region of the Horseshoe Model is marked by the subjective feeling of heightened self-control (as well as extreme rigidity that in turn hampers counterfactual self-control) is bolstered by findings such as these.

It is my hope that the foregoing discussion has effectively advocated for a socalled horseshoe model that encompasses both the pathological loss of control across mental disorder types as well as ostensibly non-pathological human action. Although this is an endeavor that certainly requires further development and discussion, I believe the empirical and philosophical factors considered above merit the adoption of this model over either a standard "spectrum" model or any other sort of account that focuses exclusively on addiction and akrasia without considering the other "half" of the model. Finally, I hope this proposal has shed light on the theoretical danger of focusing exclusively on relatable and commonplace phenomenological observations when theorizing about addiction and action theory generally, given that the present account has provided one example in which this method entirely leaves out one-half of the relevant phenomenon.

ALIENATION AND IDENTIFICATION IN PATHOLOGICAL LOSS OF CONTROL

0. Introduction

Much of our philosophical understanding of the loss of control that accompanies addiction and other mental disorders is tied in some substantial way to the *phenomenology* of a loss of control. This is understandable, given the fact that most (though not all) philosophers are not intimately acquainted with what it is like to live with a disorder of agency. Furthermore, the experiences of individuals living with mental disorders are often heterogeneous and difficult to parse philosophically. Given these confounding factors, it is easy to see why someone wanting to develop a philosophical understanding of the loss of control that accompanies addiction and other mental ailments might want to lean on the phenomenology that accompanies nonpathological instances of akrasia and then extrapolate from there.

Elsewhere (in Paper 2), I discussed one instance of this phenomenon in the interdisciplinary literature on addiction, namely in Nick Heather's (2016a, 2016b, 2020) work. Although I agreed with Heather in his insistence that there is no definite boundary between addictive and non-addictive behavior, I also disagreed with him and agreed with his interlocutor, Edmund Henden (2013, 2016), that the factors that constitute *loss of control* in addiction cannot be adequately reduced to ordinary akrasia with some added intensity. I then went on to propose a "horseshoe model" of loss of control that highlighted the deep similarities between disorders associated with opposing phenomenologies: the phenomenology of extreme self-control (e.g. anorexia nervosa) and the phenomenology of

extreme loss of control (e.g. substance use disorder). I concluded by remarking how choosing to focus on instances of nonpathological akrasia, as Heather does, puts one in a position to entirely miss the significance of the compulsive half of the horseshoe when theorizing about pathological loss of control. Once the counterfactual understanding of a lack of the capacity for self-control capacity is on the table, however, we can finally see that the focus on the link between *akrasia* and the *phenomenology of a loss of control* is a red herring for what is truly relevant with regard to pathological loss of control.

This paper takes this red herring as its starting point and tries to shed light on *why* one extreme end of the horseshoe model is standardly accompanied by the phenomenology of loss of control when the other is not. In contrast to the empiricallyheavy methodology utilized in Paper 2, however, this paper will begin firmly in the "armchair" realm of what we might call classical action theory. In particular, I will focus on a critique of the concept of externality as it relates to Harry Frankfurt's understanding of alienation, akrasia, and the self. Given that it is Frankfurt who popularized the phrase "unwilling addict" that has remained so relevant in this area of literature, I believe it is fitting to examine what is arguably one of the theoretical precursors to contemporary understandings of addiction as akrasia as exemplified by Heather. By going back to one of the sources of the commonplace contemporary understanding of akrasia and the phenomenology of loss of control, the hope is that we can begin to unpack why this problematic conflation is so often made, and why the phenomenological asymmetry of the horseshoe model exists in the first place.

This paper will begin by providing an overview of the aforementioned critique of

Frankfurt's conception of externality and alienation as provided by Tim Schroeder and Nomy Arpaly (1999). Schroeder and Arpaly's project runs in tandem to my own in many ways in that they are critical of what they view as a conflation in Frankfurt's work between akrasia and the experiences that often tend to accompany it. In this way, these authors can be seen as advancing a complementary line of argument against the red herring of analyzing akrasia too inextricably alongside the phenomenology that often accompanies akrasia. After elucidating this argument, I expand on the authors' conclusions about the relationship between externality, alienation, and the self in order to suggest a way of interpreting these concepts in light of my horseshoe model. The resultant picture offers a fuller understanding of why focusing on akrasia delivers an impoverished theory of loss of control in addiction, whereas a counterfactual understanding of self-control as articulated by the horseshoe model can make sense of the complex interplay between the sense of self and one's experience of behaviors that are in fact lacking in self-control capacity

1. Schroeder and Arpaly on Frankfurt's conception of externality

Let us begin by considering the so-called "unwilling addict" as described by Harry Frankfurt (1971, 1977). An addict of this sort is one who wishes she were not an addict but who is unsuccessful in resisting the pull of her desire to use her drug of choice. In this way the unwilling addict acts akratically, which for Frankfurt involves acting on a desire for which she has a second-order desire not to have. One need not accept Frankfurt's entire account of the structure of the will, however, in order to appreciate the initial plausibility of his description of the addict's dilemma. The key phenomenon

of interest in the case of the unwilling addict is what Frankfurt refers to as the *externality* of the offending desire.

In dissecting the theoretical constraints of Frankfurt's unwilling addict, Schroeder and Arpaly argue that the unpleasant phenomenology that has been standardly linked to the straining of one's will is not, in fact, directly connected to akrasia or to any structure of the will[22]. In effect, they claim that this phenomenology has been mislabeled and is to be properly understood as the phenomenology of being alienated from one's desires. Although they do not use the term by name, I will argue that what Schroeder and Arpaly go on to describe as the uncomfortable felt tension between one's occurrent desire and one's deep-seated self-image fits nicely with the description of cognitive dissonance in the psychology literature.

In order to orient ourselves to Schroeder and Arpaly's discussion of Frankfurt, let us return to Frankfurt's unwilling addict[23]. As I have previously stated, the unwilling addict wishes she were not an addict, and she tries in vain to abstain from her drug habit. In the end, so the story goes, her will buckles under the force of her addictive desire. According to the authors, Frankfurt's vignette contains two phenomena that Frankfurt himself often

[22] Although Schroeder and Arpaly are focused on critiquing Frankfurt's theory, the conceptual distinctions they make are taken to be applicable well beyond his work. Since their criticisms are primarily aimed at the generally intuitive cases Frankfurt offers, their conclusions will apply to any theory that adopts a similar conception of externality being intimately tied to akrasia. Since this association is precisely my target, my use of Frankfurt as a stalking horse will similarly not restrict the significance of my conclusions.

[23] Schroeder and Arpaly focus their discussion on Frankfurt's (1977) "Identification and Externality", since this is the work in which they feel Frankfurt addresses his theory of external desires most directly.

fails to keep separate. First, there is the addict's *akrasia*: her best judgment recommends that she not administer the drug, but in the end, she acts contrary to this judgment. Secondly, there is what they describe as her *alienation* from her desire to use the drug: she experiences the desire as not truly *hers*, as something akin to a foreign invader. Schroeder and Arpaly highlight the fact that these two phenomena are conceptually distinct and might plausibly come apart, thereby distancing the phenomenon of alienation from the Frankfurtian structure of the will.

At the core of the authors' criticism is their claim that Frankfurt has misidentified the pre-theoretical phenomena that he in turn uses to construct his theories. In particular, they suggest that Frankfurt's conception of externality is fleshed out in a way that implicitly links it to akrasia, thereby connecting it to the structure of the will. This is because Frankfurt chooses to label the desires such as that of the unwilling addict described above as external desires. In "Freedom of the Will and the Concept of a Person" (1971), Frankfurt analyzes externality in terms of it being a property that can be had by certain desires. In particular, a desire is an external desire just in case it is a desire that the agent prefers not to be her will (i.e., a desire for which she has a second-order desire not to act upon). Here it is clear that his understanding of externality is one that is unavoidably tied up with his understanding of the structure of the will.

In his later work, such as in "Identification and Externality" (1977), Frankfurt tweaks his understanding of externality very slightly. In its later iteration, we are to understand an external desire as a desire for which an agent's acting upon it would

necessarily render the action akratic. According to either the early or the later formulations, then, Frankfurt's concept of externality is tightly connected to his concept of akrasia in one way or another. Given that akrasia is to be understood as the phenomenon of an agent acting against her better judgment, this is unsurprising given the definition of externality that equates an external desire with a desire the agent wants not to act upon. If an agent acts upon a desire she does not want to act upon, it will presumably almost always be the case that she judged it would be best not to act upon it.

Frankfurt goes on to clarify, however, that it is possible to come to accept a vice as truly one's own while simultaneously preferring that one not have that desire. In this case, the desire would no longer be external. Here Frankfurt is envisioning cases in which an agent is resigned to the fact that she has a particular vice, and while she may prefer that she not have it, she is nonetheless accepting of the vice as truly her own. This refinement seems to be a good one, since it is easy to imagine a case in which an agent ruefully thinks to herself, "I wish I weren't so miserly and had agreed to donate to the charity, but that's *just the way I am*." In these cases, then, an agent can act akratically under Frankfurt's revised picture even if the offending desire is not external.

Although this revision does appear to make Frankfurt's picture more plausible, Schroeder and Arpaly argue that it has the unintended consequence of threatening to undermine Frankfurt's original account of externality. Recall that we started with a conception of externality in which a desire is external just in case the agent prefers that it not be her will. In the standard case, then, acting on an external desire will be an action

which is necessarily akratic. Frankfurt's added caveat, however, allows for akratic action without externality for cases in which the agent accepts the vice as truly her own. Given the initial conception of externality, however, it is unclear why a desire's *being perceived as truly one's own* should have this effect. The authors inquire, "[b]ut why are the vices which we accept as truly our own not perceived, pre-theoretically, as external, as akin to the desire of the unwilling addict?" (p. 375). They go on to suggest that "[a] plausible answer is that when acting on such vices we experience akrasia *but no alienation*" (Ibid., emphasis in original). Frankfurt, however, does not reach this same conclusion. Instead, he continues to identify external desires as desires that we have decided should play no part in our decision making. This analysis of externality, however, does not seem to do justice to the intuitive cases he describes.

It would appear, then, that Frankfurt erred when he insisted upon tying externality to akrasia and thus to the structure of one's will. One could of course respond on behalf of Frankfurt that externality is a term of art within his theory, thereby making it the case that externality means whatever he intended it to mean. Schroeder and Arpaly are right to point out, however, that Frankfurt's project gains traction by appealing to cases in which the relevant phenomena are meant to be intuitively recognizable. The case of the unwilling addict owes its intuitive appeal in large part due our ability to recognize the phenomenon of externality in the context of addictive desires. There is clearly something unique about the character of the unwilling addict's addictive desire, but the source and essence of this uniqueness need not conform with what Frankfurt himself claims, nor does it require an

association with Frankfurt's structure of the will in order to maintain its theoretical appeal. As Schroeder and Arpaly describe it, the unwilling addict's external desire is "a desire crying out for understanding" (p. 372). According to the authors, the phenomenon of externality is best analyzed not as something relating to akrasia but as a form of alienation.

In motivating this claim, the authors ask us to consider the various examples of externality Frankfurt cites. As we have already seen, the unwilling addict's desire is meant to be an external desire. In addition, Frankfurt includes intrusive, obsessional thoughts within his picture of externality. The inclusion of external thoughts sits rather awkwardly, however, with an account that analyzes externality in terms of akrasia and thus its relation to the structure of one's will. After all, it is very dubious that intrusive thoughts could count as an action flowing from the structure of the will as Frankfurt describes it. This awkwardness can be avoided, however, if we take up Schroeder and Arpaly's suggestion that a thought or a desire can count as external just in case one *feels alienated from it*.

The authors do not take these considerations to be decisive against Frankfurt's depiction of externality, however, and so they go on to offer Frankfurtian-style cases designed to highlight the conceptual distinction between akrasia and externality-as-alienation. Since this distinction is essential to what I will go on to say about the phenomenology of alienation in relation to the two extreme ends of the horseshoe model, it is worth going over the particular cases they discuss. The first case they offer is one in which the subject, Emma, feels alienated from the desire that becomes her will (in Frankfurt's terminology), and yet is clearly *not* akratic in her action. Emma, at her doctor's

suggestion, has started exercising after a lifetime of being sedentary. After a few weeks of transitioning to an active lifestyle, her friend John offers her a ride home when he notices that she appears to be experiencing some muscle pain. Emma declines John's offer, much to her surprise and to John's. When he asks if she is sure, she confirms that she is. After John leaves, Emma thinks to herself how surprised she is at her own decision to take the hard route and walk home.

Crucially, Emma acted in accordance with her own best judgment. Her surprise and mild feeling of discomfort is due to the fact that Emma still views this new, healthier version of herself as somewhat alien and unfamiliar. After all, the old Emma would have never turned down the ride home. The experience of realizing that we are different now from how we were can be a somewhat uncomfortable experience, even if we judge the change to be a positive change. Schroeder and Arpaly's case of Emma thereby shows that one can feel alienated from a desire without exhibiting anything like akrasia.

The second case the authors offer is the converse of Emma's case, i.e., a case in which the subject acts akratically but is not alienated from his desire. The case involves William, a devout Christian who experiences strong urges to spank his children when he is angry with them. Unlike Frankfurt's example of the subject who is resigned to his vices, William continues to struggle against this vice of violence whenever he is faced with it. When the urge does arise, he offers up the prayer, "Thy will, not mine, Lord" (p. 378). The authors describe the case in this way in order to suggest that William views his sinful desire to spank as more his own than his patience, since the instances in which he acts out of patience are considered to be at least partially due to the grace of God. It is important to

keep in mind that the authors need not claim that *all* cases similar to William's involve a subject who sees his sinful desires as truly his own, and his virtuous desires as being given to him by God. For their purposes, they need only establish that the particular case of William is intuitively recognizable. Given that this *does* seem to be an intuitively plausible case, Schroeder and Arpaly have provided an instance in which the subject acts akratically and yet no alienation is present.

The authors contend that the case of William contains all the theoretically significant features that would lead Frankfurt to label his desire to spank his children as an "outlaw" desire. However, William is not in any way alienated from his desire, and so there does not seem to be any reason to label his desire as external. Conversely, the case of Emma seems to clearly be non-akratic, although she is, in fact, alienated from her desire to walk. In summarizing the theoretical takeaway from their cases, they write,

> *[T]he phenomenon which unifies Frankfurt's angry man, unwilling addict, and obsessive thought examples, the phenomenon at which he seems to be gesturing and of which he provides a theory, the phenomenon which appears also to be found in Emma's case but not in William's, is alienation… What is really of interest is that the most salient unifying phenomenon, intuitively, is not some Frankfurtian structure of the will but a particular sort of experience (p. 379).*

By appealing to the same style of cases that Frankfurt himself employs, Schroeder and Arpaly have managed to divorce the salient feature of the unwilling addict, namely alienation, from akrasia. Furthermore, a commonsense notion of self-control (i.e., *not* the notion of counterfactual self-control argued for in the horseshoe model) can very plausibly be analyzed as the opposite of akrasia in these contexts, since resisting akrasia would involve controlling oneself in such a way that one's best judgment succeeds in becoming

one's will. In addition to driving a wedge between alienation and akrasia, then, the authors have also managed to conceptually distance the experience of alienation from the commonsense notion of self-control.

Why do we feel alienated from certain desires and not others, if the answer is not to be found in the structure of our wills? Schroeder and Arpaly offer the sketch of an answer that ties the unpleasant experience of alienation to the conflict between an occurrent desire and the way in which the agent views herself. In particular, they propose that this conflict must be between the offending desire and one's deep-seated self-image. They write, "[a]lienation, we would like to suggest, is *the unpleasant experience of oneself as being other than one takes oneself to be*" (p. 381, emphasis in original). They go on to clarify that this phenomenon cannot merely be caused by surprise at one's own psychological states—there must necessarily be a negative affect present. However, as in the case of Emma, this negative experience need not be due to the fact that one views the alienated desire (in her case, the desire to exercise) as negative in and of itself. Rather, in Emma's case the unpleasantness is due to the fact that significant personal changes can be somewhat threatening at first, even when they are judged to be for the best. The key ingredient for the experience of alienation, then, is a matter of *conflict* between one's self-image and the desire in question.

In this way, Schroeder and Arpaly manage to conceptually divorce the phenomenon of external desires from the structure of the will, instead opting to identify externality with the experience of alienation. This allows them to provide a sketch of what the

phenomenology at issue, namely alienation, actually consists in. Although they hypothesize that the experience of alienation is caused by a felt conflict between an occurrent desire and one's self-image, they have comparatively little to say when it comes to fleshing out this suggestion. Indeed, in giving their account of alienation in terms of a conflict with one's (typically subconscious) self-image, Schroeder and Arpaly admit that they are "treading on perilously empirical ground" (p. 382). They go on to clarify that it would suffice for their purposes "if somewhere in the psyche there exists a self-image not identical to what one is disposed to predict about oneself, a 'visceral self-image', as it might be called, which produces feelings of alienation as described" (Ibid.).

I would like to suggest that the authors need not have been so wary, since the empirical overlap they acknowledge in fact lends substantial credibility to their suggestion upon further investigation. After first introducing two new agents that I will use to elucidate my claims, I will bolster the authors' proposal with some salient empirical considerations in the following section.

2. Alienation as a conflict with the self-image and its relation to the horseshoe model

To begin, allow me to first introduce two new agents, Debbie and Anna. Debbie is an individual who is addicted to heroin in the mold of Frankfurt's unwilling addict. She has found some partial success with recovery at various points in her history, but she is currently in the midst of another relapse. She is alienated from her desire to use heroin, a desire that she experiences as external and oppressive. Next, consider Anna. Anna is an

anorexic who is in what I have elsewhere defined (Cf. Paper 1) as the "pre-awareness stage" of anorexia nervosa, which roughly amounts to exhibiting a sufficient level of anosognosia with respect to her illness[24]. An anorexic in this stage of illness is subject to what I have analyzed elsewhere as an illusion in the sense of agency, which to say that Anna experiences herself as utilizing self-control when engaging in food restriction when this is not, in fact, the case.

As I established in Paper 2, by the time Debbie and Anna have progressed to a point at which they merit their respective diagnoses, they will have undergone the neurological and psychological changes associated with disorder-related attentional biases. In addition, their respective disordered rituals will likely serve as an important means of self-regulation in their day-to-day lives, and many of these behaviors will have taken on elements of pathological habit formation. Anna, however, is unlike Debbie in that she does not feel alienated from the pathological desires relating to her disorder. In fact, the difference is even starker than a mere lack of alienation—Anna, in the grips of her disorder in its pre-awareness stage, wholeheartedly *identifies with* her anorexic desires and behaviors.

With these two cases in hand, I will turn now to the first of the empirical considerations that can lend support to an account of alienation along the lines of Schroeder

[24] As part of the DSM-5 (2013) criteria for anorexia nervosa, individuals who merit the diagnosis must exhibit "a persistent lack of recognition" with respect to the seriousness of their low body weight and restrictive behaviors. This hallmark symptom of anorexia is often described as anosognosia with respect to anorexia nervosa. Anosognosia, Greek for "to not know a disease", is a medical term first coined in order to describe stroke victims who are genuinely incapable of acknowledging that they have become paralyzed, although it has since been adopted as a term applied to other conditions such as AN.

and Arpaly. Recall that what is needed is a way to add some legitimacy to a theory that appeals to a deep-seated self-image that can either cohere with or clash with an agent's occurrent desire. I propose that one useful conceptual distinction to employ for this task is the egosyntonic-egodystonic distinction utilized in psychopathology. Although this terminology is admittedly Freudian in heritage, it continues to be useful in distinguishing otherwise symptomatically similar behaviors across differing mental disorders.

In this domain, a disorder or a behavior related to one's disorder is labeled as "egosyntonic" when it is experienced as congruent with the individual's self-image, goals, and values (Rosenthal 2003). Conversely, a disorder or behavior is labeled as "egodystonic" when the disorder or behavior in question conflicts with one's standing self-image (Ibid.). Crucially, anorexia nervosa is considered to be one of the hallmark instances of an egosyntonic disorder (Gregertsen et al. 2017). The egosyntonic label is generally reserved for AN and the personality disorders, making other psychiatric disorders (including substance use disorder) egodystonic (for a review of the egosyntonic label as applied personality disorders, see Hart et al. 2018).

The presence of the egosyntonic-egodystonic distinction in psychopathology is relevant in that it serves as one strand of empirical support that vindicates an appeal to a deep-seated self-image when distinguishing between forms of mental disorder. To be sure, this evidence is far from decisive, but it does go some way toward alleviating the worry that an appeal to an implicit and deeply held self-image is something that is empirically untenable. This distinction is made possible, after all, by positing the existence of an

implicit self-image with some form of content such that one's desires and behavior can either align with or conflict with it.

An especially useful illustration of the egosyntonic-egodystonic distinction in action can be found in the difference between obsessive compulsive disorder (OCD) and obsessive compulsive personality disorder (OCPD). OCD and OCPD are disorders with symptoms that are externally very similar: both conditions are marked by ritualistic and obsessive behavior and a fixation with control (Marchesi et al. 2008). Individuals suffering from OCD, however, experience their disorder egodystonically—in other words, they are *alienated* from their compulsions, which they see as oppressive and distressing. In contrast to OCD, OCPD is an egosyntonic disorder, which means that individuals with OCPD view their obsessive and control-oriented tendencies as reasonable and even desirable. While there is high comorbidity between AN and both OCD and OCPD, it is worth noting that the presence of OCPD within the anorexic population has been linked to especially poor treatment outcomes (Crane et al. 2007).

With the egosyntonic-egodystonic distinction in hand, we can begin to flesh out the phenomenon of alienation in this context. Anorexics, much like individuals with OCPD, will tend to derive satisfaction and even fulfillment upon completion of their disordered behaviors. This is because agents such as Anna are in the curious position of experiencing their actions that amount to a counterfactual lack of self-control in a way that is egosyntonic. By contrast, unwilling addicts such as Debbie will not experience their continued addictive behaviors as being reflective of their core values or identities. Their disordered behaviors are egodystonic, which goes hand in hand with the phenomenology

of a loss of control when drug-seeking actions are engaged in, as opposed to the phenomenology of self-control that accompanies Anna's disordered actions. These distinctions are what form the basis for placing individuals such as Anna on the leftmost "compulsive" end of the horseshoe model and individuals such as

Debbie on the rightmost "impulsive" end of the model. Relating this to Schroeder and Arpaly's understanding of alienation, to be alienated from a desire is to experience it as egodystonic, and suffering from a full-blown disorder constituted by such desires would place one on the rightmost end of the horseshoe. Similarly, a lack of alienation plus a feeling of identification/approval toward a desire would make the given desire egosyntonic, and a disorder constituted by disordered desires such as these would place one on the leftmost end of the horseshoe model.

Recall that the phenomenology of being alienated from one's (egodystonic) desire must necessarily involve a negative affect. It was then suggested that the source of this negatively-valenced experience *just is* the contradiction between one's self-image and the desire one is considering acting upon. As a second line of empirical evidence in support of this proposal, I would like to suggest that the phenomenon the authors describe fits nicely with the psychological work on dissonance theory. The research program on cognitive dissonance that is most applicable to the present discussion is the theory of dissonance related to the self-concept originally developed by Eliot Aronson (1969). This theory posits a negatively-valenced feeling of dissonance when cognitions relating to some behavior clash with the self-concept (Aronson 1992). Aronson writes that "most people hold

standards for their own behavior that are largely in accord with the conventional morals and prevailing values of society" (p. 592). He goes on to clarify that, "[s]pecifically, dissonance reduction will typically involve an effort to maintain two important elements of the self-concept: the sense of self as both (a) morally good and (b) competent" (Ibid.). This theory of dissonance has proven very successful in providing explanations for otherwise unaccounted-for experimental results[25].

Again, while not decisive, the presence and explanatory success of the cognitive dissonance literature related to the self-concept adds further plausibility to Schroeder and Arpaly's suggestion that the negative phenomenology of alienation is simply "*the unpleasant experience of oneself as being other than one takes oneself to be*" (p. 381, emphasis in original). In addition, the present proposal offers us a further piece of insight that may help to explain *why* an addict such as Debbie would experience cognitive dissonance in the midst of her addiction, whereas an anorexic such as Anna would not. From a societal perspective, drug abuse typically carries negative connotations tied up with moral failing and, indeed, poor willpower or self-control. Given this, it is unsurprising that Debbie would experience cognitive dissonance when in the grips of her addiction to heroin,

[25] Cf. Quilty-Dunn (unpublished 2019, 2020) for a philosophical treatment of rationalization and the socalled "psychological immune system" as it relates to the negative affect associated with cognitive dissonance. Quilty-Dunn's understanding of cognitive dissonance is also inextricably tied to the individual's latent self-image. Crucially, the self-image that is operant here is strikingly similar to the description of a deep-seated self-image suggested by Schroeder and Arpaly that is not rationally based and is overly optimistic in the contents it ascribes to the subject.

if we assume that her self-image, like that of most people, carries the assumptions that she is *good* and *competent*[26].

In contrast, the values of self-control, perfectionism, and persistence are firmly established as positive attributes in our society and are even frequently tied to moral praiseworthiness (think of all of the historical saints with so-called "holy anorexia"!). Since these are attributes that anorexics like Anna tend to hold especially dear, attributes that they view their disordered behavior as exemplifying, it is easy to see how cognitive dissonance would be absent. Given that Anna is anosognosic in her pre-awareness stage and is thereby still ignorant of the fact that her asceticism has morphed into something pathological, she will not be alienated from her form of disordered desire. As a result, Anna does not experience the phenomenology of straining that Debbie does experience, the phenomenology that was incorrectly associated with the Frankfurtian understanding of akrasia.

At this point it will hopefully be less mysterious why agents such as Anna, who occupy the leftmost "compulsive" end of the horseshoe, do not fit into the tidy theoretical box of experiencing something like akrasia when they engage in their pathological behaviors. I view Schroeder and Arpaly's contributions as providing invaluable insight into one of the very influential ways (owing to Frankfurt) in which philosophers have often

[26] To be clear, I am *not* claiming that individuals with substance abuse disorders are not, in fact, good and competent. Indeed, I am not even necessarily claiming that Debbie, when asked, would agree with such statements. It is a common psychological phenomenon, after all, for individuals to internalize negative societal tropes or beliefs about a social category to which they belong without having consciously accepted the content of such beliefs—consider, e.g., internalized misogyny in women and the various forms of impostor syndrome.

conflated the phenomenon of genuinely lacking self-control and the unpleasant *experience* that often accompanies such scenarios. Instead, it was revealed that the difference between Anna and Debbie (who, recall, occupy similarly extreme ends on the horseshoe model and thus have very little counterfactual capacity for self-control) is a matter of the congruency between their pathological desires and their self-images. In this way, choosing to focus on the *negative affect* of cognitive dissonance present in the case of Debbie but not Anna obscures the way in which both agents lack self-control in a strikingly similar fashion, as argued for in Paper 2.

The way forward from here, I would like to suggest, does not lie in any attempt to resuscitate the idea of ordinary akrasia and its accompanying phenomenology as a philosophical starting point for theorizing about loss of control in mental disorder. The foregoing has hopefully shown that this route leads to an overly narrow and simplified version of what is in fact a rich and complicated interplay between one's disordered actions and one's self-image. I believe the way forward must involve a deeper exploration into how our self-concepts influence our own experiences and subsequent interpretations of our actions. It is unsurprising that the deep similarities between anorexia nervosa and drug addiction have been overlooked for so long, since certain assumptions regarding addiction, self-control, and the role of the self-concept have long been underdeveloped, and indeed AN itself is vastly underdeveloped in the philosophical literature. Now that these connections and distinctions have been made, the way forward must involve a careful examination of the feedback loop between our self-image and our perception of our own

agency, which will in turn allow us to get closer to the root of loss of control in its pathological and non-pathological forms.

References

Ainslie, George (2001). *Breakdown of Will*. Cambridge: Cambridge University Press.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. Arlington, VA.

Aronson, Eliot

--(1969). "The theory of cognitive dissonance: A current perspective". *Advances in Experimental Social Psychology*, 4: 1-34.

--(1992). "The return of the repressed: Dissonance theory makes a comeback". *Psychological Inquiry*, 3(4): 303-311.

Attia, Evelyn. (2010). "Anorexia nervosa: current status and future directions". *Annual Review of Medicine*, 61:425–435.

Bayne, Tim and Pacherie, Elisabeth (2007). "Narrators and comparators: The architecture of agentive self-awareness". *Synthese*, 159: 475-491.

Bayne, Tim

--(2008). "The Phenomenology of Agency". *Philosophy Compass*, 3(1): 182-202.

--(2011). "The Sense of Agency", in *The Senses*, ed. Fiona Macpherson. Oxford: Oxford University Press.

Bickel et al. (2014). "The Behavioral Economics of Substance Use Disorders: Reinforcement Pathologies and Their Repair". *Annual Review of Clinical Psychology*, 10: 641-677.

Bickel, W.K. and Li, R. (2010). "Neuroeconomics of addiction: the contribution of executive dysfunction", in Ross et al. (eds.), *What is Addiction?*, MIT Press:

Cambridge, MS, p. 1-26.

Bickel, Warren and Marsch, Lisa (2001). "Toward a behavioral economic understanding of drug dependence: Delay discounting processes". *Addiction*, 96: 73-86.

Birgegard et al. (2009). "Anorexic Self-Control and Bulimic Self-Hate: Differential Outcome Prediction from Initial Self-Image". *International Journal of Eating Disorders*, 42(6): 522-530.

Blakemore, Sarah-Jayne and Frith, Chris (2003). "Self-Awareness and Action". *Current Opinion in Neurobiology*, 13: 219-224.

Brooks, Samantha (2016). "A debate on working memory and cognitive control: can we learn about the treatment of substance use disorders from the neural correlates of anorexia nervosa?". *BioMed Central Psychiatry*, 16:10.

Brooks et al.

--(2012). "A debate on current eating disorder diagnoses in light of neurobiological findings: is it time for a spectrum model?". *Biomed Central Psychiatry*, 12:76.

--(2017). "The Role of Working Memory for Cognitive Control in Anorexia Nervosa versus Substance Use Disorder". *Frontiers in Psychology*, 8:1651.

Bulik et al. (2006). "Prevalence, Heritability, and Prospective Risk Factors for Anorexia Nervosa". *Archives of General Psychiatry*, 63(3): 305-312.

Butlin, Patrick and Papineau, David (2016). "Normal and addictive desires", in Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. Oxford: Oxford University Press.

Case et al. (2012). "Diminished size-weight illusion in anorexia nervosa: Evidence for

visuo-proprioceptive integration deficit". *Experimental Brain Research* 217(1): 79-87.

Charland et al. (2013). "Anorexia Nervosa as a Passion". *Philosophy, Psychiatry, & Psychiatry*, 20(4): 353-365.

Coniglio et al. (2017). "Won't stop or can't stop? Food restriction as a habitual behavior among individuals with anorexia nervosa or atypical anorexia nervosa". *Eating Behaviors*, 26: 144-147.

Cox et al. (2006). "The addiction-Stroop test: Theoretical considerations and procedural recommendations". *Psychological Bulletin*, 132(3): 443-476.

Crane et al. (2007). "Are obsessive-compulsive personality traits associated with a poor outcome in anorexia nervosa? A systematic review of randomized controlled trials and naturalistic outcome studies". *International Journal of Eating Disorders*, 40(7): 581588.

Davidson, Donald

-- (1980). "How is weakness of the will possible?", in Donald Davidson (ed.), *Essays on Actions and Events*. Oxford: Clarendon Press, p. 21-42.

--(1982). "Paradoxes of irrationality", in R. Wollheim, J. Hopkins (eds.), *Philosophical Essays on Freud*. Cambridge University Press: Cambridge, UK, p. 289-305.

Dill, Brendan and Holton, Richard (2014). "The addict in us all". *Frontiers in Psychiatry*, 5(139): 1-20.

Draper, Heather (2000). "Anorexia Nervosa and Respecting a Refusal of Life Prolonging

Therapy: A Limited Justification. *Bioethics*, 14:120-133.

Eshkevari et al. (2012). "Increased plasticity of the bodily self in eating disorders". *Psychological Medicine*, 42(4): 819-828.

Fairburn et al. (1998). "A cognitive-behavioral theory of anorexia nervosa". *Behavioral Research and Therapy*, 37:1-13.

Flanagan, Owen (2013). "Willing addicts? Drinkers, dandies, druggies, and other Dionysians", in Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. Oxford: Oxford University Press, p. 66-81.

Frank et al. (2012). "Anorexia nervosa and obesity are associated with opposite brain response". *Neuropsychopharmacology*, 37(9): 2031-2046.

Frankfurt, Harry

--(1971). "Freedom of the Will and the Concept of a Person". *Journal of Philosophy*, 68: 5-20.

--(1977). "Identification and Externality", in Amelie Rorty (ed.), *The Identities of Persons*. University of California Press.

Frith, Chris (1992). *The Cognitive Neuropsychology of Schizophrenia*. East Sussex: Lawrence Erlbaum Associates Ltd.

Giordano, Simona (2005). *Understanding Eating Disorders: Conceptual and Ethical Issues in the Treatment of Anorexia and Bulimia Nervosa*. Oxford: Oxford University Press.

Graybiel, Ann. (2008). "Habits, rituals, and the evaluative brain". *Annual Review of Neuroscience*, 31: 359-387.

Gregertsen et al. (2017). “The Egosyntonic Nature of Anorexia: An Impediment to Recovery in Anorexia Nervosa Treatment”. *Frontiers in Psychology*, 8: 2273.

Hart et al. (2018). “Are personality disorder traits ego-syntonic or ego-dystonic? Revisiting the issue by considering functionality”. *Journal of Research in Personality*, 76: 124-128.

Heather, Nick

--(1998). “A conceptual framework for explaining drug addiction”. *Journal of Psychopharmacology*, 12: 3-7.

--(2016a). “On defining addiction”, in Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. Oxford: Oxford University Press, p. 3-28.

--(2016b). “Addiction as a form of akrasia”, in Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. Oxford: Oxford University Press, p. 133-152.

--(2020). “The concept of akrasia as the foundation for a dual systems theory of addiction”. *Behavioural Brain Research*, 390: 112666.

Henden, Edmund

--(2013). Addictive actions. *Philosophical Psychology* 26(3): 362-382.

--(2016). “Addiction, compulsion, and weakness of will: a dual-process perspective”, in Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. Oxford: Oxford University Press, p. 116-132.

Holton, Richard and Berridge, Kent (2013). "Addiction between compulsion and choice", in Neil Levy (ed.) *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. Oxford: Oxford University Press, p. 239-268.

Holton, Richard (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Hope et al. (2013). "Agency, ambivalence and authenticity: The many ways in which anorexia nervosa can affect autonomy". *International Journal of Law in Context*, 9(1):20-36.

Hornbacher, Marya (1999). *Wasted: a Memoir of Anorexia and Bulimia*. New York, NY: Harper Perennial.

Inzlicht, Michael and Schmeichel, Brandon (2012). "What is ego depletion? Toward a mechanistic revision of the resource model of self-control". *Perspectives on Psychological Science*, 7(5): 450-463.

Jacquemot, Aimée Margarita Marisol Catherine, and Park, Rebecca (2020). "The Role of Interoception in the Pathogenesis and Treatment of Anorexia Nervosa: A Narrative Review". *Frontiers in Psychiatry*, 11(98): 1-8.

Kahneman, Daniel (2011). *Thinking, Fast and Slow* (vol. 1). New York: Holt.

Kaplan et al. (2009). "The slippery slope: prediction of successful weight maintenance in anorexia nervosa". *Psychological Medicine*, 39:1037–1045.

Kaye et al. (2009). "New insights into symptoms and neurocircuit function of anorexia nervosa". *Nature Reviews Neuroscience*, 10(8): 573-584.

Keating, Charlotte (2010). "Theoretical perspective on anorexia nervosa: The conflict of

reward". *Neuroscience and Biobehavioral Reviews*, 34(1): 73-79.

Keating et al. (2012). "Reward processing in anorexia nervosa". *Neuropsychologia*, 50: 567-575.

Keys, Ansel (1950). *The Biology of Human Starvation*. Minneapolis: University of Minnesota Press. Moore, James W. (2016). "What Is The Sense of Agency and Why Does it Matter?".

King et al. (2019). "Cognitive Overcontrol as a Trait Marker in Anorexia Nervosa? Aberrant Task- and Response-Set Switching in Remitted Patients". *Journal of Abnormal Psychology*, 128(8): 806-812.

Kurzban, Robert (2010). "Does the brain consume additional glucose during self-control tasks?". *Evolutionary Psychology*, 8(2): 244-259.

Marchesi et al. (2008). "Temperament features in adolescents with ego-syntonic or egodystonic obsessive compulsive symptoms". *European Child & Adolescent Psychiatry*, 17: 392-396.

McKay, Ryan and Efferson, Charles (2010). "The subtleties of error management". *Evolution and Human Behavior*, 5(31): 301-319.

Mele, A.R.

-- (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press.

--(2012). *Backsliding: Understanding Weakness of Will*. Oxford: Oxford University Press

Moore, James W. (2016). "What Is The Sense of Agency and Why Does it Matter?". *Frontiers in Psychology*, 7, 1272.

Mortimer, Rose (2015). “More than just a label: identity, diagnosis and recovery from eating disorders”. PhD dissertation, King’s College London, Department of Social Science, Health and Medicine.

Muraven, Mark and Baumeister, Roy (2000). “Self-regulation and depletion of limited resources: Does self-control resemble a muscle?”. *Psychological Bulletin,* 126(2): 247-259.

Mylopoulos, Myrto

--(2014). “Agentive awareness is not sensory awareness”. *Philosophical Studies,* 169(2): 761-780.

--(2017). “A cognitive account of agentive awareness”. *Mind & Language,* 32: 545-563.

Naccache et al. (2005). “Effortless control: executive attention and conscious feeling of mental effort are dissociable”. *Neuropsychologia*, 43(9): 1318-1328.

Nordbo et al. (2006). “The Meaning of Self-Starvation: Qualitative Study of Patients’ Perception of Anorexia Nervosa”. *International Journal of Eating Disorders*, 39(7): 556-564.

O’Hara et al. (2015). “A reward-centered model of anorexia nervosa: A focused narrative review of the neurological and psychophysiological literature”. *Neuroscience and Biobehavioral Reviews*, 52: 131-152.

Olatunji et al. (2010). “Mediation of symptom changes during inpatient treatment for eating disorders: the role of obsessive-compulsive features”. *Journal of Psychiatric Research* 44(14): 910-916.

Osler, Lucy (2020). “Controlling the noise: a phenomenological account of Anorexia

Nervosa and the threatening body". *Philosophy, Psychiatry, and Psychology*.

Pacherie, Elisabeth

--(2008). "The phenomenology of action: A conceptual framework". *Cognition*, 107(1): 179-217.

--(2010). "Self-Agency", in S. Gallagher (ed.), *The Oxford Handbook of the Self*. Oxford: Oxford University Press.

Papadopoulos et al. (2009). "Excess mortality causes of death and prognostic factors in anorexia nervosa". *British Journal of Psychiatry* 194(1): 10-17.

Papezova et al. (2005). "Elevated pain threshold in eating disorders: Physiological and psychological factors". *Journal of Psychiatric Research*, 39: 431-438.

Park et al. (2014). "Hungry for reward: How can neuroscience inform the development of treatment for Anorexia Nervosa?". *Behavior Research and Therapy* 62: 47-59.

Pickard, Hanna (2017). "Addiction", in Timpe et al. (eds.), *The Routledge Companion to Free Will*. Routledge: New York, p. 454-467.

Pinto et al. (2014). "Capacity to Delay Reward Differentiates Obsessive Compulsive Disorder and Obsessive Compulsive Personality Disorder". *Biological Psychiatry*, 75(8): 653-659.

Pollatos et al. (2008). "Reduced perception of bodily signals in anorexia nervosa". *Eating Behaviors* 9(4): 381-388.

Quilty-Dunn, Jake (2020). "Unconscious Rationalization, or: How (Not) to Think about Awfulness and Death". *Unpublished Manuscript*

Riemer et al. (2013). "Action and Perception in the Rubber Hand Illusion".

Experimental Brain Research, 224(3): 383-393.

Robinson, Terry and Berridge, Kent (1993). "The neural basis of drug craving: An incentive-sensitization theory of addiction". *Brain Research: Brain Research Reviews* 18(3): 247-291.

Rosenthal, Howard (2003). *Human Services Dictionary* p. 102. Brunner-Routledge.

Schroeder, Timothy and Arpaly, Nomy (1999). "Alienation and Externality". *Canadian Journal of Philosophy*, 29(3): 371-387.

Schwabe, Lars and Wolf, Oliver (2009). "Stress Prompts Habit Behavior in Humans". *Journal of Neuroscience* 29(22): 7191-7198.

Setiya, Kieran (2007). *Reasons Without Rationalism*. Princeton, NJ: Princeton University Press.

Skog, Ole- Jørgen (2003). "Rational capacities, or: How to distinguish recklessness, weakness, and compulsion", in Sarah Stroud and Christine Tappolet (eds.), *Weakness of Will and Practical Irrationality*. New York: Oxford University Press, p. 17-38.

Steinglass et al.

--(2006). "Set shifting deficit in anorexia nervosa". *Journal of the International Neuropsychological Society*, 12: 431-435.

--(2012). "Increased capacity to delay reward in anorexia nervosa". *Journal of the International Neuropsychological Society*, 18: 773-780.

--(2018). "Targeting habits in anorexia nervosa: a proof-of-concept randomized trial". *Psychological Medicine*, 48(15): 2584-2591.

Steinhausen, Hans-Christoph

--(2002). “The outcome of anorexia nervosa in the 20th century”. *American Journal of Psychiatry*, 159:1284–1293.

--(2009). “Outcome of Eating Disorders”. *Child and Adolescent Psychiatric Clinics of North America*, 18(1): 225-242.

Szasz, Thomas (1974). *The Myth of Mental Illness*: Foundations of a Theory of Personal Conduct. New York: Harper & Row.

Szmukler, George and Tantam, Digby (1984). “Anorexia nervosa: Starvation dependence”. *British Journal of Medical Psychology*, 57: 303-310.

Tchanturia et al. (2012). “Poor cognitive flexibility in eating disorders: Examining the evidence using the Wisconsin Card Sorting Task”. *PLoS ONE*, 7: e28331.

Vandereycken, Walter (2006). “Denial of Illness in Anorexia Nervosa—A Conceptual Review: Part 2, Different Forms and Meanings”. *European Eating Disorders Review*, 14: 352-368.

Vitousek et al. (1998). “Enhancing Motivation for Change in Treatment-Resistant Eating Disorders”. *Clinical Psychology Review*, 18(4): 391-420.

Walsh, Timothy (2013). “The Enigmatic Persistence of Anorexia Nervosa”. *American Journal of Psychiatry*, 170: 477-484.

Warin, Megan (2004). “Primitivizing Anorexia: The Irresistible Spectacle of Not Eating”. *The Australian Journal of Anthropology*, 15(1): 95-104.

Wiers et al. (2016). “Passion’s slave? Conscious and unconscious processes in alcohol and drug abuse”, in K. Sher (ed.), *Oxford Handbook of Substance Use and Substance Use Disorders*, OUP: Oxford, p. 1-80.