Anorexia nervosa: Illusion in the sense of agency

Preprint draft—please cite published version.

Amanda Evans

University of Antwerp

Abstract

This paper first identifies and then provides a novel analysis of a feature of anorexia nervosa (AN) that has been largely overlooked in the philosophy of psychiatry literature. This feature is the discrepancy between first-personal experiences of anorexic food restriction and the clinical descriptions of these same behaviors at the level of agentive awareness. I develop a positive account of the sense of agency in AN 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.

> *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*)

1. 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.

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[1] 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 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.

---

[1] 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 pertaining to 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 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".

The paper will proceed as follows. In Section 1, I will introduce anorexia nervosa from the perspective of anorexics' reports. Then, in Section 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 Section 1. Finally, in Section 3, I will resolve the tension between the descriptions of AN discussed in Sections 1-2 by offering a positive account of how the sense of agency in anorexia nervosa comes to be illusory.

2. 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 are ostensibly 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 (Warin, 2004, 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 from the times after which anorexics have come to realize they are not in control. The latter category of experiences will be covered in Section 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.[2] 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, 2004, 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

[2] 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 necessary conditions for achieving some level of recovery.

source of pride and meaning. In the following section, I will provide an 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.

3. 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 come to realize 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 nervosa 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 they 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 research. 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.[3]

[3] 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 purpose here is to show that the empirical literature

The first theory, which I will refer to as the "Habit Model" of anorexia nervosa, attempts to explain why AN 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.

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 that are insensitive to devaluation. 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) (Walsh, 2013, 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, it is insensitive to

---

and the anorexic's phenomenological testimony are *prima facie* at odds with one another but can ultimately be rendered compatible.

outcome devaluation, 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 be consciously aware of the behavior while performing the relevant action—in other words, this need not be something along the lines of an automatic reflex. Rather, a habit in this sense is a behavior that is no longer performed in the service of achieving the initial reward, since the behavior has become so overtrained that it is insensitive to outcome devaluation. Importantly, this subset of behavior does not involve the same level of planning, effort, or willfulness as action-outcome (i.e. goal-directed) actions.

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 this paper, 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.[4] 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.

[4] 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.

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,

> 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 (Hope et al., 2013, p. 24).

If either or both 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 Section 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 apparent tension will be the task of the remaining sections.

4. 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 a 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 is 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.

4.1 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 associated with 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.[5] 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 in this subfield 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,[6] 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 to the second as the "comparator" approach.

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

[5] 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" fits well with what I am describing.

[6] 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 of agentive awareness.

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.

4.2 Egosyntonicity and the solution to the puzzle

Bringing the focus back to anorexia nervosa, 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 Section 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 Section 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 intended to explain how we come to gain awareness and insight into our own 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.[7] 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 goal-directed 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,

[7] 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.

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 agentive judgments throughout the day as they engage in disordered behavior, I must admit that I do not see any clear way to adjudicate between 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 occurring will 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.[8]

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 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

[8] 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).

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 associated with 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.[9] 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 at which 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 willpower once they have already become encoded as habit. In the anorexic case, however, this is precisely what is occurring.

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 relevant pre-awareness experiences in AN and the Habit and Reward Models' descriptions is pre-theoretic in that it is already anticipated in the reports of individuals who are no longer in the pre-awareness stage of the disorder. 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

[9] 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 a personal level, it is unlikely that cue-driven behaviors are experienced as the direct result of conscious effort and willfulness.

> 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 (Charland, 2013, 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 that this account will help to shed light on the mechanisms underpinning anorexia nervosa that remain poorly understood in both the empirical and philosophical literature. Furthermore, I hope it 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.

ACKNOWLEDGMENTS

I owe many thanks to Michelle Montague, Galen Strawson, Tim Bayne, and David Sosa for extensive feedback on multiple drafts of this paper. Thank you also to Meredith McFadden for helpful comments at an APA session on this project and to several anonymous reviewers, one of whom provided especially thorough and insightful feedback on more than one occasion.

REFERENCES

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

Bayne, T., & Pacherie, E. (2007). Narrators and comparators: The architecture of agentive self awareness. *Synthese, 159*, 475–491.

Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass*, *3*(1), 182–202.

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

Blakemore, S. and Frith, C. (2003). Self-awareness and action. *Current Opinion in Neurobiology*, 13, 219-224.

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

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.

Draper, H. (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.

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

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

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

Graybiel, A. (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.

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, M. (1999). *Wasted: A Memoir of Anorexia and Bulimia*. New York, NY: Harper Perennial.

Jacquemot, A., and Park, R. (2020). The role of interoception in the pathogenesis and treatment of anorexia nervosa: A narrative review. *Frontiers in Psychiatry*, 11(98), 1-8.

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.

Keys, A. (1950). *The Biology of Human Starvation*. Minneapolis: University of Minnesota

Press.
Moore, J. (2016). What is the sense of agency and why does it matter?". *Frontiers in Psychology*, 7, 1272.
Mortimer, R. (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.
Mylopoulos, M. (2014). Agentive awareness is not sensory awareness. *Philosophical Studies*, 169(2), 761-780.
Mylopoulos, M. (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.
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, L. (2020). Controlling the noise: a phenomenological account of anorexia nervosa and the threatening body. *Philosophy, Psychiatry, and Psychology*, 28(1), 41-58.
Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179-217.
Pacherie, E. (2010). Self-agency. In Gallagher, S., (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.
Pollatos et al. (2008). Reduced perception of bodily signals in anorexia nervosa. *Eating Behaviors* 9(4), 381-388.
Riemer et al. (2013). Action and Perception in the Rubber Hand Illusion. *Experimental Brain Research*, 224(3), 383-393.
Robinson, T. and Berridge, K. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research: Brain Research Reviews* 18(3), 247-291.
Schwabe, L. and Wolf, O. (2009). Stress Prompts Habit Behavior in Humans. *Journal of Neuroscience* 29(22), 7191-7198.
Steinglass et al. (2018). Targeting habits in anorexia nervosa: a proof-of-concept randomized trial. *Psychological Medicine*, 48(15), 2584-2591.
Steinhausen, H. (2002). The outcome of anorexia nervosa in the 20th century. *American Journal of Psychiatry*, 159,1284–1293.
Steinhausen, H. (2009). Outcome of eating disorders. *Child and Adolescent Psychiatric Clinics of North America*, 18(1), 225-242.

Szmukler, G. and Tantam, D. (1984). Anorexia nervosa: Starvation dependence". *British Journal of Medical Psychology*, 57, 303-310.

Vandereycken, W. (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, T. (2013). The enigmatic persistence of anorexia nervosa. *American Journal of Psychiatry*, 170, 477-484.

Warin, M. (2004). Primitivizing anorexia: The irresistible spectacle of not eating. *The Australian Journal of Anthropology*, 15(1), 95-104.