Mental Disorders Involve Limits on Control, Not Extreme Preferences

Chandra Sripada

Published in *Agency in Mental Disorder: Philosophical Dimensions*,
Matt King & Joshua May (eds.), Oxford University Press (2022).

1. Introduction

People with mental illness engage in characteristic disorder-associated behaviors. A person with obsessive compulsive disorder (OCD) washes their hands dozens or hundreds of times a day. A person with attention-deficit/hyperactivity disorder (ADHD) is distractible and disorganized and fails to complete their assigned tasks. A person with alcoholism drinks to excess, with resulting harms to work and family. How are we to make sense of why these people do what they do?

A standard position is that those with mental illness cannot help but do what they do. They have a disorder and what they do is not a matter of choice. We would not blame a person with acromegaly for having too much growth hormone; so too we should not blame a person with ADHD for distractedly forgetting to go to an appointment.

There are two major shortcomings of this simple "disease model" of mental illness. First, it seems to require two different models to explain action. Most purposive actions are explained in the usual way in terms of the ordinary workings of our motivational psychology—beliefs, desires, deliberation, etc. Some purposive actions, the disorder-associated actions of those with mental illness, get explained in a quite different way. For these special cases, a "disease-based" process is invoked, though the particulars of how this process works are not filled in with any detail. Splitting up explanations in this way, especially without providing details on how the second kind of explanation is supposed to work, seems *ad hoc*. Second, while the person with acromegaly has no ability whatsoever to (directly) control their growth hormone level, not so for the person with a mental disorder. For example, if one were to put a gun to the head to the person with OCD, they would straightaway desist from washing their hands.

Observations such as these have fueled an alternative perspective that sees mental illness not as a disease, but as a matter of purpose and choice. This "volitional" view has a long history. It is visible in Foucault, Laing, and Szasz (Szasz 1997; Foucault 1988; Laing 1960). It is also seen in newer critiques by Gene Heyman, Hannah Pickard, and Carl Hart (Heyman 2010; Pickard 2012; Hart 2014). The economist Bryan Caplan offers a particularly clear articulation of this volitional position.[1] Using key ideas from consumer theory, Caplan distinguishes constraints *on* actions from preferences *for* actions. He argues physical illnesses produce constraints on one's actions. Mental illnesses do not; they are best understood in terms of volition, albeit in the context of extreme preferences that are out of step with societal norms.

My aim in this chapter is to offer a systematic response to the volitional view of mental illness. The core of my argument is that theorists who support the volitional view operate with a too simple model of human motivational architecture. They view the human mind as having a decision theoretic structure: We have various desires, they differ in strength (reflecting strength of preference), and we always do what we

[1] Caplan, Bryan. 2006. "The Economics of Szasz: Preferences, Constraints and Mental Illness." Rationality and Society 18 (3): 333–66. My interest in Caplan's article was spurred by Scott Alexander's discussion on the Slate Star Codex Blog (https://slatestarcodex.com/2020/01/15/contra-contra-contra-caplan-on-psych/).

most prefer. I argue the human mind instead has a regulatory control structure. We not only have desires (or similar spontaneous states; I use the term "desire" to refer to all these states for the time being), we have regulatory mechanisms that enable us to modulate or suppress our desires. The presence of regulatory mechanisms introduces the possibility of constraints: if regulation is limited in some way, then certain "lesser" desires that do not reflect what we most prefer may still manifest in action. This is in fact what happens, I argue, in many mental disorders—these disorders arise precisely where the limits of control are breached (in interestingly different ways in different disorders). If this picture is right, then a person's disorder-associated behaviors might not reflect what they most prefer to do, but rather what they are constrained to do.

This chapter is divided into three parts. Section 2 distinguishes two models of motivational architecture, the Decision Theory model and the Regulatory Control model. Section 3 adopts the Regulatory Control model and sketches a general picture of several major mental illnesses. They are, I argue, conditions that arise due to limits on control. Section 3 returns to the key distinction between preferences and constraints. It is argued that the limits on control explanation of mental illness is a better overall fit to the data than the volitional view.

2. Two Models of Motivational Architecture

2.1. The Decision Theory Model

There is a picture of motivational architecture that is extremely common in philosophy, economics, and certain social sciences. The picture resembles a psychologized version of rational choice theory, and it goes like this: People have various desires directed at different things. These desires differ in terms of *strength* (Mele 1998). That is, there are certain motivational properties of these desires in virtue of which they are ordered in terms of motivational "potency" (barring ties—I ignore this complication going forward). Action selection systems are configured so that they are sensitive to the strength properties of one's overall set of desires, and the desire that sits atop the strength ordering becomes the basis for action. The explanations for action supplied by this model are simple and intuitive, for example: Joe's desire to go to the movies is stronger than his desire to do anything else (go to the park or go to the mall, etc.), and so Joe goes to the movies.

This picture of motivational architecture, which I will call the Decision Theory view is so widespread in philosophy and economics, it hardly gets noticed or mentioned. It simply serves as the background default view for understanding agents. But the view implies two principles that are worth pausing to highlight.

First, because of the way the Decision Theory architecture links one's strongest desire to action, the architecture implies that what an agent does will conform to the following law-like generalization, which has been dubbed the Law of Desire:

> Whenever a person acts intentionally, they do what they are most strongly motivated to do at the time.[2]

The second principle, which is a direct consequence of the first, is what we can call the Law of Revealed Desire:

[2] See Mele 2003; Sripada 2014; Barnes 2019 for discussion. The principle as stated is susceptible to counterexamples but these counterexamples are not relevant for our present purposes. Thus, I prefer this simpler formulation for the present purposes.

Whenever a person acts intentionally, what they do reveals what is their strongest motive.

That is, when a Decision Theory agent acts, we can "read off" from their behavior what they most want. Extending this second principle to mental illness, the volitional view naturally follows. Suppose someone touches a doorknob and washes their hands a dozen times in a row, and now they are washing their hands (intentionally) for the thirteenth time, causing serious skin fissures and substantial pain. The second principle implies that washing their hands this thirteenth time is what they most wanted. To be sure, they have unusual wants, or what Caplan calls "extreme preferences". But, if we assume a Decision Theory architecture correctly describes human motivation, then since this is what they intentionally do, we can be confident this is what they most wanted to do.

2.2. Regulatory Control Model

I now turn to an alternative picture of motivational architecture. As we go about ordinary life, various kinds of *spontaneous tendencies* arise. Our attention is grabbed by features of the environment. Habitual action tendencies are elicited. Memory items are spontaneously called to mind. We are "pulled" to think about certain topics. A hallmark of spontaneous tendencies such as these is that they operate as a default—the spontaneous tendencies will manifest in action unless something intervenes to block them.[3]

Such intervention is possible because humans have unique abilities for top-down regulation. In what follows, I discuss these regulatory abilities in two steps. First, I discuss regulation of simple, brief spontaneous tendencies of the kind just considered. Second, I discuss regulation of more complex, temporally-extended states such as emotions and cravings.

The regulation of simple, brief spontaneous tendencies is called *cognitive control*, and it is extensively studied in cognitive and clinical neuroscience. A standard method involves study of "conflict tasks". The hallmark of these tasks is that they set up a conflict between the simple spontaneous tendencies previously discussed and a second type of motivational state, one's *goals*. These are relatively stable motivational states that are closely connected to one's conscious reflective judgments. Here are three examples of conflict tasks:

> *Stroop Task* (Stroop 1935) – On each trial, subjects are shown a color word ("red", "blue") which is itself printed in an ink color. Subjects are asked to state the ink color of the word on all trials. On congruent trials, the word's meaning and ink color match and it is relatively easy to get the right answer. On incongruent trials, the word's meaning and ink color are discrepant, and subjects must exert control over their spontaneous tendency to read the word, in order to select the correct response.
>
> *Go/No Go Task* (Donders 1969) – On each trial, subjects see a letter on the screen. Subjects are asked to press a button only if the letter is not "X" and withhold the button press if it an "X". Most of the letters are not "X", for example 90% not "X" to 10% "X". This skewed ratio leads to the development of a habit for button pressing. On trials where the stimulus is not "X", the button pressing habit facilitates correct responding. On "X" trials, subjects must suppress this habit.
>
> *Think/No Think Task* (Anderson and Green 2001) – During a practice session, subjects are trained to recall pairs of words (e.g., ROACH – ORDEAL; GUM – TRAIN). In the test session, they are given the first member of the pair. They are told that if the word appears in green ink, they are to think about

[3] In Sripada (forthcoming), I discuss the nature of these simple spontaneous tendencies in some detail.

the paired word. If the word appears in red ink, they must not think about the paired word. This requires that they suppress the spontaneous tendency to recall the associated word.

These tasks illustrate that based on their goals, people can perform *control actions*—rapidly executed intra-psychic actions that inhibit, suppress, or otherwise modulate various kinds of simple, brief spontaneous tendencies. Different kinds of control actions target different psychological systems. As a result, people can control a diverse array of simple, spontaneous tendencies, including those associated with attention, memory, thought, belief formation, evaluation, and action selection.

Turn now to complex, "hot", temporally-extended spontaneous states such as emotions and cravings. We can regulate these states as well in accordance with our goals. Theorists call this capacity various names including "effortful control", "volitional regulation", and "emotion regulation" (Gross 1998; Rothbart et al. 2003; Sripada et al. 2014). This kind of regulation is illustrated vividly in fMRI studies of craving regulation (Brody et al. 2007; Kober et al. 2010; Hare, Camerer, and Rangel 2009). In these studies, subjects, for example smokers or dieters, are shown pictures of stimuli (cigarettes, indulgent food, etc.) that are known to elicit strong cravings. On some trials, they are asked to simply experience the cravings. On other trials, they are asked to regulate the cravings and reduce their intensity. This is usually accomplished by attention control actions (directing attention away from pictures) and thought control actions (intentionally inhibiting certain thoughts or bringing to mind competing thoughts). These studies typically find:

1) elevated activation in reward-related regions during experience trials;
2) elevated activation in "executive" regions during regulation trials; and
3) an inverse relationship between activity in executive regions and reward regions (suggesting the former is inhibiting the latter).

It is an interesting question how regulation of complex spontaneous states such as emotions and cravings relates to cognitive control, i.e., regulation of the simple, brief spontaneous tendencies I discussed earlier. I discuss this issue in detail elsewhere (Sripada forthcoming). In short, I think the two are related as whole and part: When a person regulates complex states, they perform a sequence of cognitive control actions directed at simple, brief spontaneous tendencies. I put this issue aside for our present purposes.

I need a general term to refer to this broad collection of spontaneous states, either simple or complex, irrespective of whether they pertain to belief, memory, thought, or action selection. I refer to them all as "pulses". I also need a term to describe different forms of goal-directed regulation, spanning cognitive control over simple, brief states and more complex forms of regulation over complex states. Going forward, I refer to them all as "regulatory control", or "regulation" for short.

Recall the two principles that characterize the Decision Theory agent, the Law of Desire and the Law of Revealed Desire. Critically, they need not hold in a Regulatory Control agent if one additional condition is met: regulation is limited. If regulation is in some way inefficient, weak, or fallible, then an agent can do things that they themselves do not most want to do[4], in violation of the Law of Desire. This will happen when three conditions hold:

1) A person's strongest overall desire is to do one thing (e.g., pay attention to a lecture),
2) They experience spontaneous pulses to do something else (e.g., to notice the ticking of the clock or to mind wander onto some meaningless topic).

[4] Importantly, what an agent "most wants" is determined by motivational properties of desires, not by observing which desire actually manifests in action. See Mele 2003; Sripada 2014 for further discussion.

3) Top-down regulation is in some way limited, allowing spontaneous pulses to manifest in action (recall that pulses are motivational defaults and thus will be the basis for action unless they are regulated).

When these conditions hold, the person will act on pulses rather than on what they most want, in violation of the Law of Desire. It follows that what they do also fails to reflect what they most want, in violation of the Law of Revealed Desire.

Now, the details here are complex because with a Regulatory Control agent, there are different, somewhat independent sources of motivation arising from the states that I have been calling goals and pulses. Thus, the notion of "what an agent most wants" is more challenging to define: Is it one's strongest goal? One's strongest pulse? Can their respective strengths even be compared? I do not want to get bogged down in these details.[5] It suffices for our purposes to take note of the fact that with a Regulatory Control agent, even if they have the sincere goal of doing one thing, due to limitations on regulatory control, they can still end up doing something else.

2.3. Humans have a Regulatory Control Architecture

There is extensive evidence, reviewed elsewhere (Botvinick and Cohen 2014; Cohen 2017; Hofmann, Schmeichel, and Baddeley 2012), that human motivational architecture has a regulatory control structure, and that is what I will be assuming going forward. Notice, though, that even if the Regulatory Control model is correct, the Decision Theory model remains useful. For example, in most ordinary contexts where there is no need for regulation, or where regulation is so easy it operates flawlessly, then the Decision Theory model and Regulatory Control model will yield similar behavioral predictions. So, the Decision Theory model represents a simplification that works fairly well for day-to-day purposes. However, and this is critical, the two models do come apart in some contexts, and mental illness, as I will presently argue, is a striking example.

3. Mental Illness and Dyscontrol

Having introduced the Regulatory Control model of motivational architecture, I now want to fill in the details of how, given this architecture, regulatory control fails in ways relevant to mental disorders. I refer to a state in which the limits of control are breached as a *dyscontrol state*. At a highly general level, all dyscontrol states arise from a mismatch between regulation efficacy and pulse efficacy, which in turn arises from one of three possibilities: an elevated "load" of pulses, a decrease in the person's regulatory capacities, or both. Once we move past this generalization, however, and look at specific mental disorders, we find a variety of types of mismatches that are operative. These mismatches involve different types of pulse states (e.g., attentional, emotional, doxastic), different types of decreases or impairments in regulatory capacities, and different types of environmental contexts in which the pulse/regulation mismatches unfold. I will discuss four disorders to illustrate this variety. Along the way, I highlight certain Control Limiting Factors that arise in these disorders that illuminate specific and interestingly different pathways by which the limits of control are breached. An important theme that emerges in what follows is that the Control Limiting Factors that are relevant to psychiatric disorders involve "extended limits"—limits observable only across days, months, and even years.

[5] I discuss this issue in some detail in Sripada (2014).

3.1. Obsessive Compulsive Disorder (OCD)

Individuals with OCD have obsessive thoughts, which are typically directed at characteristic themes (e.g., contamination), and these thoughts arouse substantial anxiety and tension. They additionally have repeated urges to perform behaviors related to these obsessive thoughts, for example urges to wash their hands. Importantly, these thoughts and urges do not just happen occasionally, for example a few times a week or several times a day. Rather, they typically occur with much greater frequency: dozens to hundreds of times a day, often occupying a significant portion of the day.

Consider an individual OCD thought (e.g., the thought that one's hand is contaminated) or an individual OCD urge (e.g., the urge to wash one's hand). Each one of these is readily susceptible to regulation. The person can use the regulatory repertoire discussed in the previous section to redirect attention, suppress problematic thoughts, inhibit inappropriate action tendencies, and so on. Regulation will tend to fail, however, if a person experiences densely recurrent thoughts and urges throughout the day, day after day, month after month. Under these circumstances, regulation starts to become too burdensome for the person.[6]

One kind of burden is experiential. The exercise of top-down regulatory capacities is associated with a distinctive effortful phenomenology that is aversive or otherwise negatively valenced (Shenhav et al. 2017). Thus, spending significant stretches of one's day engaging in top-down regulation of thoughts and urges burdens the person with prolonged dysphoric feelings.

A second kind of burden arises from opportunity cost. Top-down regulation is a member of a larger set of cognitive functions called executive functions (Diamond 2013). Other members include planning, deliberation, and high-level problem-solving. Executive functions are underpinned by a shared, or importantly overlapping, set of brain mechanisms that exhibit limited capacity—engaging these executive mechanisms for one purpose entails, for the most part, giving up their use for other purposes.[7] It follows that if a person must engage in top-down regulation for significant stretches of their day, they must pay substantial opportunity costs in foregoing a range of other valuable executive activities—planning, deliberation, problem-solving—that they could have otherwise undertaken.

In short then, the Control Limiting Factor that operates in OCD involves *cumulative burden*. No thought or urge in the disorder is, by itself, particularly hard to control. But when we consider them in their temporal totality—that is, when we consider the cumulative burden of having to regulate all of these densely recurrent thoughts and urges over extended stretches of time, the burden on the person is excessive and regulation predictably falters.

The obsessional thoughts and urges in OCD illustrate a more general phenomenon that I claim is found in most mental disorders. We see in OCD three key features: 1) a massive population of pulse-type states; 2) the pulses are in some recognizable sense abnormal; 3) the presence of these abnormal pulses is a long-term feature of the person's psychology. Going forward, I refer to this cluster with a convenient short-hand name: "CAPPs", for *chronic aberrant populations of pulses*. Giving the phenomenon a name will, it is hoped, make it easier to recognize just how ubiquitous it is across a wide range of psychiatric disorders.

2.2. Attention-Deficity/Hyperactivity Disorder (ADHD)

[6] I discuss burdens of regulation in OCD in Sripada (forthcoming).

[7] There is substantial evidence, especially from neuroimaging, of a single domain-general executive network (Duncan and Owen 2000; Niendam et al. 2012; Cole and Schneider 2007). The limited capacity of this network is supported by a number of lines of evidence, see Baddeley 1996; Kurzban et al. 2013 for partial reviews.

In ADHD, we see CAPPs, but rather than obsessive thoughts and impulses, the CAPPs pertain to attention.[8] As we transact with the environment, features of the environment "call out" for our attention (Corbetta and Shulman 2002; Serences et al. 2005): a whisper in the hallway, the text message that may have shown up on one's phone, one's own internal spontaneous musings and mind wanderings. This is true for all individuals, with or without ADHD—attentional pulses impinge on the psyche day in and day out.

Most of these attentional pulses do not present much of a problem because we can regulate them, thus staying on task and avoiding inappropriate distraction. Moreover, unlike OCD, regulating attentional distractors is not particularly effortful or dysphoric, and thus it does not create a cumulative burden on the person. In ADHD, however, a problem arises because there is a regulation/pulse mismatch: either attentional pulses are too frequent or regulation efficacy is diminished[9], leading to a higher than typical failure rate in which inappropriate attentional pulses more frequently "get through". To be clear, individuals with ADHD *can* regulate attentional pulses, and indeed they succeed most of the time. The problem they face is instead statistical. The modern world is unforgiving in placing demands on our attention; tasks and projects at school and at work require unerring focus to get done well, or get done at all. A higher error rate in regulating attentional distractors is enough to create mistakes, forgetfulness, and disorganization—the core symptoms of ADHD. The main Control Limiting Factor that is operative in ADHD is *fallibility*. The point probability of successfully regulating each attentional pulse remains quite high. But because attentional pulses are so ubiquitous, the person still experiences regular errors, which in turn produce serious negative academic, occupational, and interpersonal consequences.[10]

3.3. Major Depressive Disorder

The hallmark of major depression is the presence of the emotion sadness—not mild sadness that is temporary, but severe sadness that is persistent (i.e., on most occasions for an extended duration). Emotions such as sadness produce a multitude of effects on one's psychology that are mediated by pulse-type states, as I have argued elsewhere in detail (Sripada forthcoming). Here are some of sadness's effects (Freed and Mann 2007; Hybels et al. 2009; Cipriani et al. forthcoming; Gaddy and Ingram 2014): one's attentional patterns are changed—negative or potentially threatening features of the environment now spontaneously draw one's attention. One's spontaneous interpretations change—ambiguous or neutral events are now interpreted in a negative or pessimistic light. One's thoughts change—negative memories about the past or pessimistic prospections about the future spontaneously enter one's mind. One's action tendencies change—there is a pervasive sense of fatigue that makes doing even basic things feel overwhelmingly effortful. In short, then, depression is a condition that involves chronic alterations in pulses arising from multiple psychological systems; that is, it involves CAPPs.

As in the other cases, these pulses associated with attention, belief, thought, or action can be regulated. For example, a person can suppress a negative memory or force themselves to get out of bed on any particular occasion if supplied with sufficient incentives. The relevant question, however, is whether in real world circumstances where such salient incentives are absent and these problematic pulses arise nearly continuously, can an ordinary person without specialized training in higher-order control regulate them? The answer is "no" and for multiple reasons. One factor is *deference*. Ordinary people's default position is to

[8] I am focusing here on ADHD, inattentive type, the most common type in adults with ADHD. A broadly similar account could be given of ADHD hyperactive type and ADHD impulsive type, where the role of attentional pulses is replaced by motoric pulses and reward-seeking/appetitive pulses, respectively.

[9] There are few attempts to distinguish which of these two factors predominates in ADHD (cf., Friedman-Hill et al. 2010). However, at least some individuals with ADHD have more wide-ranging difficulties with executive functions suggesting that for them, the top-down factor is more heavily implicated.

[10] I discuss fallibility in the context of cumulative risk of relapse in addiction at length elsewhere (Sripada 2018).

accept their spontaneously-formed beliefs and impressions. It is rare for people to take a meta-cognitive stance and check carefully whether the way things seem corresponds to the way things actually are. A second factor is *vigilance failure*. Without specialized training in sustained meta-cognitive monitoring, an ordinary person cannot stand at guard monitoring and regulating their own ongoing beliefs, impressions, and thoughts continuously. A third factor is *lack of regulatory skill*. Suppose a person does manage to recognize that something is "off" about an impression that arises on a particular occasion— say, the impression that nobody likes them. Simple suppression strategies might succeed in pushing the thought out of their mind for a moment, but thoughts such as these often immediately return.

Now, there are more sophisticated ways to defeat such thoughts. For example, cognitive behavioral therapy (Aaron T Beck 1963; 1964; Aaron T. Beck 1979) trains a person to systematically challenge the evidential basis of problematic automatic thoughts, so that undermining these thoughts becomes routinized and more permanent. Advanced meditative training seeks to impart comprehensive control over how attention is directed and how thoughts arise (Rubia 2009). Skills such as these, however, are an *achievement*; they are attained by relatively few, and they are not something that ordinary people simply execute as a matter of course. *Deference*, *vigilance failure*, and *lack of regulatory skill* might each be considered Control Limiting Factors take alone. When they operate together, they surely constitute limits on one's control.

3.4. Schizophrenia

In schizophrenia, we once again see the operation of CAPPs. According to a leading theory, the central cognitive/motivational alteration in schizophrenia is abnormal salience.[11] *Salience* refers to a property of a stimulus to grab attention and become the target of valenced appraisal. In schizophrenia, ordinary day-to-day stimuli acquire inappropriate hypertrophied salience: a smile by a stranger, two people coincidentally sharing the same name, a dog with a distinctive limp. These events are passed over in neurotypical individuals, but in individuals with schizophrenia, they strike the person as deeply important and self-relevant, and they become the targets of spontaneous interpretative activity to try to make sense of them. Over time (typically years), ongoing interpretive activity targeting countless events and situations crystalizes in the formation of a delusional system, a system of internally coherent beliefs that makes sense of the person's subjective experience.

Now, for most people, the formation of odd, bizarre beliefs—ones that that are wildly out of step with one's other beliefs about the world and that are not shared with others in one's cultural milieu—are noticed by the person (De Neys and Glumicic 2008; Mercier 2020). This in turn generate efforts, mediated by executive systems (i.e., systems that implement top-down regulatory control), to challenge and correct the errant beliefs. Strikingly, this does not happen in schizophrenia. Thus, a second factor is likely at work: reduced monitoring. Ongoing surveillance of beliefs, already somewhat lax in neurotypical individuals, is compromised still further in schizophrenia.[12] Thus, errant beliefs evade executive correction processes and remain in place, and over time, they become entrenched.

In short then, schizophrenia is a disorder whose etiology is rooted in CAPPs. Abnormalities in salience lead to ongoing bombardment with "doxastic pulses": spontaneous appraisals of day-to-day events in distorted (often paranoid) ways. Many of the Control Limiting Factors discussed earlier likely play a role in explaining why these pulses are not regulated: ordinary people are excessively deferent to their

[11] Kapur 2003; Howes and Kapur 2009. The theory actually pertains to psychosis. Schizophrenia is more complex syndrome with psychosis as a central element.

[12] Dopamine dysfunction provides a unifying explanation of why the two deficits are paired: midbrain dopamine pathways are involved in salience processing (Kapur 2003) while mesocortical dopamine pathways are involved in executive functions, which include monitoring (Braver, Barch, and Cohen 1999; Goldman-Rakic et al. 2004).

spontaneous impressions; inappropriate doxastic pulses overload executive correction mechanisms; people are inexpert at challenging ill-founded beliefs. And there are likely additional factors, such as impaired monitoring of errant beliefs, that are operative in schizophrenia specifically.

3.5. Summing Up

In this section, I discussed four major psychiatric disorders. My analysis of what goes on in these disorders had a common structure: All these disorders centrally involve chronic aberrant populations of pulses or CAPPs, and, in some cases, there were also inefficiencies or impairments in regulatory capacities. In each disorder, the disorder-associated pulses *over the long-term* breach certain limits of control, thus explaining why the person exhibits the characteristic disorder-associated symptoms. Space does not allow me to discuss more disorders or conditions, such as addiction, mania, or anxiety disorders. But the general form of how I would explain these conditions is already clear. In short then, on my view, a key feature of many major mental disorders is that they involve limits on control.[13]

4. Preferences or Constraints Revisited

Are mental disorders best explained in terms of limits on control, or do they reflect volition in the setting of extreme (and socially stigmatized) preferences? I now want to do some argument "scorekeeping" comparing the two views, focusing on some of Caplan's arguments.

4.1. Incentive Sensitivity and the "Gun to the Head Test"

One of the main arguments for the volitional view involves the "gun-to-the-head-test". Caplan, for example, writes:

> Can we change a person's behavior purely by changing his incentives? If we can, it follows that the person was able to act differently all along, but preferred not to; his condition is a matter of preference, not constraint… (Caplan 2006, 349).

Here Caplan presents crisply and succinctly what is probably the most common theme in a vast "anti-psychiatry" literature: mental disorders involve choices that are stigmatized, but there is no genuine loss of control or impairments in agency. Variants of the gun-to-the-head test (or more general incentive sensitivity tests) are put forward by Pickard, Hart, Heyman, Morse, Foddy & Savalescu, and many others (Pickard 2012; Hart 2014; Heyman 2010; Morse 2002; Foddy and Savulescu 2010).

We are now in a position to see why conclusions based on the gun-to-the-head-test are misleading. In my account of mental disorders, I emphasized the role of chronic pulses, i.e., CAPPs, and I identified a number of Control Limiting Factors that arise specifically in that context, which in turn lead to the characteristic thoughts and actions we see in these disorders. The gun-to-the-head test, however, describes a scenario with little to no relevance to mental disorders because CAPPs are absent and none of the Control Limiting Factors have a chance to operate.

[13] This weaker claim is what I need for the present purposes in critiquing Caplan. I actually endorse a stronger view: There is a deep conceptual tie between mental disorder and dyscontrol, and thus a*ll* mental disorders involve limits on control. I will develop this view in due course, but I do not try to defend it here.

Consider a person with OCD. They can stop washing their hands if you put a gun to their head. But they still face limits on control that arise from the *cumulative burden* of having to regulate an unending, recurrent series of dysphoric urges. If you put a gun to the head of someone with ADHD, they can regulate a certain distracting attentional pulse (indeed they succeed at this anyways most of the time). Their problem is one of *fallibility* in the context of temporally-extended projects, and the gun-to-the-head-test has nothing to say about this. A person with depression can interpret a situation less negatively if you threaten them with certain death. But this threat is explicit and, by stipulation, definitive. In their day-to-day lives, however, they need to regulate ongoing negative interpretations and thoughts that lack this kind of clarity and certitude, allowing *deference*, *overload*, and *lack of regulatory skill*, among other Control Limiting Factors, to operate. A person with schizophrenia can be ordered under threat of serious harm to re-interpret events in less paranoid ways. But such interventions, if they work at all, are invariably temporary. Due to continued aberrant salience attribution and *impaired monitoring* of errant beliefs, among other Control Limiting Factors, spontaneous paranoid interpretations will soon return and their delusional system will be reinstated.

The gun-to-the-head test initially strikes us as plausible because we have a picture that when agency breaks down, barriers to purposive actions are decisive, rigid, and easy to see with a quick look. For example, in contrasting mental illness and physical illness, Caplan notes that no incentive can get someone who is paralyzed to stand (Caplan 2006, 342). Here the absence of incentive sensitivity is clear with a single glance. Many theorists similarly seem to assume that mental disorders need to impair agency in a similarly decisive and easy to check way. But dyscontrol in mental disorders is, as we have seen, not much like this at all. It instead involves temporally extended faltering of agency due to the cumulative impact of CAPPs, with substantial preserved incentive sensitivity at any given slice of time.

Now, to be absolutely clear, I am not arguing that, in contrast to physical disorders, mental disorders yield only weak constraints on thought and action. That is actually the opposite of my view; I believe mental disorders produce constraints on thought and action that are serious and severe. A person bombarded with obsessive thoughts of contamination and urges to hand wash is in a very real sense coerced (intra-psychically) into doing what they do. My point is that with mental illness, there is not one single or even several decisive blow that can be easily spotted, but rather countless tiny cuts that may much be harder to appreciate.

4.2. A Unified Model of Human Behavior

Another claimed advantage of the volitional view is that it presents a unified model of behavior. All purposive behavior, both healthy and disordered, is explained as arising from one's ordinary preference-based motivational psychology. Caplan complains that economists have been too willing to carve out a special exception for mental illness, as if the laws of preference-based behavior apply everywhere else but somehow not there. He writes:

> Though these authors are usually eager to bring social phenomena into the orbit of economics, they not only make an exception for severe mental illness; they treat the exception as uncontroversial. Over time, however, diagnoses of mental illness have become increasingly widespread. Epidemiologists now report that 20% or more of the USA population suffers from mental illness during a given year (Kessler et al. 1994). A seemingly small loophole in the applicability of economics has grown beyond recognition.

The limits on control view, however, avoids Caplan's charge because the view invokes a single model of motivation for all behavior, the Regulatory Control model. Most ordinary behavior arises within the

"regulation frontier" of the architecture: regulation works properly either because it is not needed (the relevant pulse states are situationally appropriate) or because it succeeds in subduing problematic pulse states. In some cases, the limits of control of this Regulatory Control architecture are systematically breached and the person regularly exhibits dyscontrol characteristic of a psychiatric illness. But there are not two models of behavior here. There is a single model that involves multiple parameters (e.g., efficacy of pulses, efficacy of regulation, etc.), and health and disease occupy different regions of the parameter space. Just like a model of a car engine explains both why a Mustang hums and why it sputters, the Regulatory Control model explains purposive agency in both health and disease.

4.3. Dystonicity

I now turn to a feature of mental illness that is hard for the volitional view to explain, but makes perfect sense with the limits on control model.

Many mental disorders are "ego dystonic": The person repudiates, rejects, or in some other way "stands against" their disorder-associated thoughts and actions (Freud 2014; Clark 1992; Belloch, Roncero, and Perpiñá 2012; Purdon et al. 2007). To be sure, some disorders are not dystonic in this way, at least overtly. For example, people with paranoid schizophrenia do not typically come to the doctor seeking out help with their delusions. But with many disorders, e.g. OCD, ADHD, and depression, people with the conditions actively seek out clinical care and pursue fairly demanding treatments. A natural explanation for why they do this is that there is something about their thoughts and actions that they dislike and want to change. But this natural interpretation makes little sense on the volitional view of mental illness.

To make this point concrete, take a person with ADHD. According to supporters of the volitional view such as Caplan, this person most prefers to chase variety and distraction—that is why they are disorganized, forgetful, and scattered. If that is truly their strongest preference—that is, if their preference ranking really is *chasing variety/distraction* > *being organized*—then it is puzzling why they are at the clinic month after month working with a psychiatrist on a medication regimen and working with a behavioral therapist on extensive cognitive/behavioral treatments.

Defenders of the volitional view might respond that we need to distinguish what the person herself prefers from societal reactions and stigma. While the person herself most prefers variety and distraction, she is nonetheless at the clinic to change her thoughts and actions because that is "what society demands". But this response falters because it relies on an inappropriately restrictive understanding of preferences. If chasing variety and distraction is tightly linked to the emergence of interpersonal problems for the person, then we need to change the descriptions of their options to reflect this. We thus assess their preferences over the following "conjoined" outcomes: *chasing variety/distraction and incurring interpersonal problems* vs. *not chasing variety/distraction and not incurring interpersonal problems*. If the person prefers the latter, then they do not have a problem according to the Decision Theory model of motivation, i.e., the model that undergirds the volitional view. They will just straightaway not chase variety and distraction and avoid the interpersonal problems that would have ensued. But if they prefer the former, then we are back to our original problem: Why are they in the clinic week after week undertaking costly and burdensome treatments to rid themselves of thoughts and actions that, according to the volitional view, they actually genuinely prefer? The volitional view does not seem to have a good answer.

The limits on control view, on the other hand, has a ready explanation for dystonicity. The person with ADHD is in the clinic week after week because she has the goal of being organized and timely and thereby achieving all the positive consequences that flow from that (occupational and interpersonal success, etc.). But she is beset by distracting attentional pulses that arise irrespective of these goals, and, though she can regulate many, or even most, of these attentional pulses, she cannot successfully regulate all of them—that is, she has reached a limit on control. So, she now finds herself doing all sorts of things—for example,

being forgetful and disorganized—that she does not really want to do. The basic form of this explanation generalizes to a wide range of psychiatric disorders. In addition to OCD and depression (discussed earlier), it extends to other conditions, such as anxiety and addiction, where, though I did not discuss them, it is not hard to see how to apply the general form of this model.

Stepping back a bit, the fundamental problem for volitional view of mental illness is that to explain dystonicity that clearly attends many mental disorders, we need a way for agents to regularly and recurrently do things that they prefer not to do, even hate to do (e.g., wash their hands for the 100th time or have lapses of attention for the thousandth time). The volitional view, however, relies on the Decision Theory model of motivation. As such, it obeys the Law of Desire, which says roughly that agents do what they most want to do. But by tying action so tightly to preference, this law seems to make dystonicity, especially chronic dystonicity of the kind seen in psychiatry, impossible.[14]

5. Conclusion

Consumer theory distinguishes between one's preferences, what one wants to do, and one's budget, what one is able to do. The volitional view of mental illness locates mental illness on the preference side—mental illness involves choice rather than constraints on what one is able to do. The choices are, to be sure, sharply out of step with societal norms and are thus stigmatized, but they remain just that: choices.

In responding to volitional view of mental illness, I put forward a more structured model of motivational architecture, one that countenances both spontaneous states as well as regulatory capacities that are responsive to our goals and that regulate these spontaneous states. But regulation has its limits, especially when it must be deployed over extended intervals of time (months and years) against massive populations of spontaneous tendencies to think and do various things. The existence of limits on control opens up space for agents to regularly and recurrently think things and do things that they themselves prefer not to think and do, and mental disorders, I argued, reside in this space. The constraints on thought and action found in mental disorders are certainly different in kind from constraints in physical conditions. They are, nonetheless, no less real.

References

Anderson, Michael C., and Collin Green. 2001. "Suppressing Unwanted Memories by Executive Control." *Nature* 410 (6826): 366.

[14] One move available to supporters of the volitional view is to appeal to "meta-preferences" (see for example Caplan's blog post "The Depression Preference" (https://www.econlib.org/the-depression-preference/). The idea is that a depressed person prefers to lie in bed and think guilty thoughts. At the same time, however, they "meta-prefer" to not have depressive first-order preferences and that is why they find their depressive behaviors ego dystonic and seek treatment. However, if we try to combine the meta-preference view with the Decision Theory model of motivation and its associated Law of Desire, the result is incoherence. The problem can be stated in the form of a dilemma. If the first-order preference to lay in bed is the person's strongest, then why is the person at the clinic week after week seeking to defeat this desire? On the other hand, if the meta-preference is the person's strongest, then the basic premise of the volitional view is falsified. The person with depression does *not* most prefer to lie in bed and think guilty thoughts as originally claimed; they actually most prefer essentially the *opposite*. That is, they strongly disprefer having the motives on the basis of which they do these things, and they want those first-order preferences to be eradicated. If we go further and try to explain why the person cannot seem to bring about what they most strongly meta-prefer, the most plausible answer is that they there is some constraint that prevents them. In this way, the meta-preference view quickly ends up abandoning volition in favor of constraints.

Baddeley, Alan. 1996. "Exploring the Central Executive." *The Quarterly Journal of Experimental Psychology Section A* 49 (1): 5–28.

Barnes, Eric Christian. 2019. "An Argument for the Law of Desire." *Theoria* 85 (4): 289–311.

Beck, Aaron T. 1963. "Thinking and Depression. I. Idiosyncratic Content and Cognitive Distortions." *Arch Gen Psychiatry* 9: 324–33.

———. 1964. "Thinking and Depression. II. Theory and Therapy." *Arch Gen Psychiatry* 10: 561–71.

Beck, Aaron T. 1979. *Cognitive Therapy of Depression*. Guilford press.

Belloch, Amparo, María Roncero, and Conxa Perpiñá. 2012. "Ego-Syntonicity and Ego-Dystonicity Associated with Upsetting Intrusive Cognitions." *Journal of Psychopathology and Behavioral Assessment* 34 (1): 94–106.

Botvinick, Matthew M., and Jonathan D. Cohen. 2014. "The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers." *Cognitive Science* 38 (6): 1249–85. https://doi.org/10.1111/cogs.12126.

Braver, Todd S., Deanna M. Barch, and Jonathan D. Cohen. 1999. "Cognition and Control in Schizophrenia: A Computational Model of Dopamine and Prefrontal Function." *Biological Psychiatry* 46 (3): 312–28.

Brody, Arthur L., Mark A. Mandelkern, Richard E. Olmstead, Jennifer Jou, Emmanuelle Tiongson, Valerie Allen, David Scheibal, Edythe D. London, John R. Monterosso, and Stephen T. Tiffany. 2007. "Neural Substrates of Resisting Craving during Cigarette Cue Exposure." *Biological Psychiatry* 62 (6): 642–51.

Caplan, Bryan. 2006. "The Economics of Szasz: Preferences, Constraints and Mental Illness." *Rationality and Society* 18 (3): 333–66.

Cipriani, Andrea, A. Tomlinson, B. Teufer, A. M. Chevance, G. Gartlehner, S. Touboul, P. Ravaud, C. Le Berre, E. I. Fried, and V. T. Tran. forthcoming. "Identifying Outcomes for Depression That Matter to Patients, Informal Caregivers and Healthcare Professionals: Qualitative Content Analysis of a Large International Online Survey." *Lancet Psychiatry*.

Clark, David A. 1992. "Depressive, Anxious and Intrusive Thoughts in Psychiatric Inpatients and Outpatients." *Behaviour Research and Therapy* 30 (2): 93–102.

Cohen, Jonathan D. 2017. "Cognitive Control." In *The Wiley Handbook of Cognitive Control*, 1–28. Wiley-Blackwell. https://doi.org/10.1002/9781118920497.ch1.

Cole, Michael W., and Walter Schneider. 2007. "The Cognitive Control Network: Integrated Cortical Regions with Dissociable Functions." *NeuroImage* 37 (1): 343–60. https://doi.org/10.1016/j.neuroimage.2007.03.071.

Corbetta, Maurizio, and Gordon L Shulman. 2002. "Control of Goal-Directed and Stimulus-Driven Attention in the Brain." *Nature Reviews. Neuroscience* 3 (3): 201–15. https://doi.org/10.1038/nrn755.

De Neys, Wim, and Tamara Glumicic. 2008. "Conflict Monitoring in Dual Process Theories of Thinking." *Cognition* 106 (3): 1248–99.

Diamond, Adele. 2013. "Executive Functions." *Annual Review of Psychology* 64: 135–68.

Donders, Franciscus Cornelis. 1969. "On the Speed of Mental Processes." *Acta Psychologica* 30: 412–31.

Duncan, John, and Adrian M Owen. 2000. "Common Regions of the Human Frontal Lobe Recruited by Diverse Cognitive Demands." *Trends in Neurosciences* 23 (10): 475–83. https://doi.org/10.1016/S0166-2236(00)01633-7.

Foddy, Bennett, and Julian Savulescu. 2010. "A Liberal Account of Addiction." *Philosophy, Psychiatry, & Psychology: PPP* 17 (1): 1.

Foucault, Michel. 1988. *Madness and Civilization: A History of Insanity in the Age of Reason*. Vintage.

Freed, Peter J., and J. John Mann. 2007. "Sadness and Loss: Toward a Neurobiopsychosocial Model." *American Journal of Psychiatry* 164 (1): 28–34.

Freud, Sigmund. 2014. *On Narcissism: An Introduction*. Read Books Ltd.

Friedman-Hill, Stacia R., Meryl R. Wagman, Saskia E. Gex, Daniel S. Pine, Ellen Leibenluft, and Leslie G. Ungerleider. 2010. "What Does Distractibility in ADHD Reveal about Mechanisms for Top-down Attentional Control?" *Cognition* 115 (1): 93–103.

Gaddy, Melinda A., and Rick E. Ingram. 2014. "A Meta-Analytic Review of Mood-Congruent Implicit Memory in Depressed Mood." *Clinical Psychology Review* 34 (5): 402–16.

Goldman-Rakic, Patricia S., Stacy A. Castner, Torgny H. Svensson, Larry J. Siever, and Graham V. Williams. 2004. "Targeting the Dopamine D 1 Receptor in Schizophrenia: Insights for Cognitive Dysfunction." *Psychopharmacology* 174 (1): 3–16.

Gross, James J. 1998. "The Emerging Field of Emotion Regulation: An Integrative Review." *Review of General Psychology* 2: 271–99.

Hare, T. A., C. F. Camerer, and A. Rangel. 2009. "Self-Control in Decision-Making Involves Modulation of the VmPFC Valuation System." *Science* 324: 646–48. https://doi.org/324/5927/646 [pii] 10.1126/science.1168450.

Hart, Carl. 2014. *High Price: A Neuroscientist's Journey of Self-Discovery That Challenges Everything You Know About Drugs and Society*. Reprint edition. New York, NY: Harper Perennial.

Heyman, Gene M. 2010. *Addiction: A Disorder of Choice*. Reprint edition. Cambridge, Mass.; London: Harvard University Press.

Hofmann, Wilhelm, Brandon J. Schmeichel, and Alan D. Baddeley. 2012. "Executive Functions and Self-Regulation." *Trends in Cognitive Sciences* 16 (3): 174–80.

Howes, Oliver D., and Shitij Kapur. 2009. "The Dopamine Hypothesis of Schizophrenia: Version III—the Final Common Pathway." *Schizophrenia Bulletin* 35 (3): 549–62.

Hybels, Celia F., Dan G. Blazer, Carl F. Pieper, Lawrence R. Landerman, and David C. Steffens. 2009. "Profiles of Depressive Symptoms in Older Adults Diagnosed with Major Depression: Latent Cluster Analysis." *The American Journal of Geriatric Psychiatry* 17 (5): 387–96.

Kapur, Shitij. 2003. "Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia." *American Journal of Psychiatry* 160 (1): 13–23.

Kober, Hedy, Peter Mende-Siedlecki, Ethan F. Kross, Jochen Weber, Walter Mischel, Carl L. Hart, and Kevin N. Ochsner. 2010. "Prefrontal–Striatal Pathway Underlies Cognitive Regulation of Craving." *Proceedings of the National Academy of Sciences* 107 (33): 14811–16. https://doi.org/10.1073/pnas.1007779107.

Kurzban, Robert, Angela Duckworth, Joseph W. Kable, and Justus Myers. 2013. "An Opportunity Cost Model of Subjective Effort and Task Performance." *The Behavioral and Brain Sciences* 36 (6): 661–79. https://doi.org/10.1017/S0140525X12003196.

Laing, R. D. 1960. *The Divided Self: An Existentialist Study in Sanity and Madness*. Penguin.

Mele, Alfred. 1998. "Motivational Strength." *Noûs* 32 (1): 23–36. https://doi.org/10.1111/0029-4624.00085.

———. 2003. *Motivation and Agency*. New York: Oxford Univerity Press.

Mercier, Hugo. 2020. *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton University Press.

Morse, Stephen J. 2002. "Uncontrollable Urges and Irrational People." *Virginia Law Review* 88: 1025–78.

Niendam, Tara A., Angela R. Laird, Kimberly L. Ray, Y. Monica Dean, David C. Glahn, and Cameron S. Carter. 2012. "Meta-Analytic Evidence for a Superordinate Cognitive Control Network Subserving Diverse Executive Functions." *Cognitive, Affective, &Behavioral Neuroscience* 12 (2): 241–68. https://doi.org/10.3758/s13415-011-0083-5.

Pickard, Hanna. 2012. "The Purpose in Chronic Addiction." *AJOB Neuroscience* 3 (2): 40–49. https://doi.org/10.1080/21507740.2012.663058.

Purdon, Christine, Emily Cripps, Matthew Faull, Stephen Joseph, and Karen Rowa. 2007. "Development of a Measure of Egodystonicity." *Journal of Cognitive Psychotherapy* 21 (3): 198–216.

Rothbart, Mary K., Lesa K. Ellis, M. Rosario Rueda, and Michael I. Posner. 2003. "Developing Mechanisms of Temperamental Effortful Control." *Journal of Personality* 71 (6): 1113–44.

Rubia, Katya. 2009. "The Neurobiology of Meditation and Its Clinical Effectiveness in Psychiatric Disorders." *Biological Psychology* 82 (1): 1–11.

Serences, John T., Sarah Shomstein, Andrew B. Leber, Xavier Golay, Howard E. Egeth, and Steven Yantis. 2005. "Coordination of Voluntary and Stimulus-Driven Attentional Control in Human Cortex." *Psychological Science* 16 (2): 114–22.

Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. "Toward a Rational and Mechanistic Account of Mental Effort." *Annual Review of Neuroscience* 40: 99–124.

Sripada, Chandra. forthcoming. "Loss of Control In Addiction: The Search For An Adequate Theory And The Case For Intellectual Humility." In *Oxford Handbook of Moral Psychology*, edited by John M. Doris and Manuel Vargas. https://umich.box.com/s/soc6hhnz9yexm09r6hg8dugd60opu0um.

———. forthcoming. "The Atoms of Self-Control." *Nous*.

———. 2014. "How Is Willpower Possible? The Puzzle of Synchronic Self-Control and the Divided Mind." *Noûs* 48: 41–74.

———. 2018. "Addiction and Fallibility." *The Journal of Philosophy* 115 (11): 569–87.

Sripada, Chandra, Michael Angstadt, Daniel Kessler, K. Luan Phan, Israel Liberzon, Gary W. Evans, Robert C. Welsh, Pilyoung Kim, and James E. Swain. 2014. "Volitional Regulation of Emotions Produces Distributed Alterations in Connectivity between Visual, Attention Control, and Default Networks." *NeuroImage* 89 (April): 110–21. https://doi.org/10.1016/j.neuroimage.2013.11.006.

Stroop, J. Ridley. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18 (6): 643.

Szasz, Thomas. 1997. *Insanity: The Idea and Its Consequences*. Syracuse University Press.