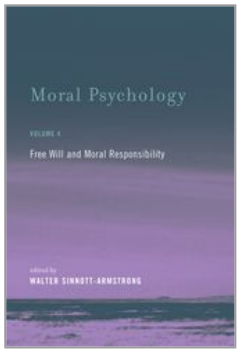
University Press Scholarship Online
MIT Press Scholarship Online



Moral Psychology, Volume 4: Free Will and Moral Responsibility
Walter Sinnott-Armstrong

Print publication date: 2014
Print ISBN-13: 9780262026680
Published to MIT Press Scholarship Online: September 2014
DOI: 10.7551/mitpress/9780262026680.001.0001

The Freedom to Choose and Drug Addiction

P. Read Montague

DOI:10.7551/mitpress/9780262026680.003.0008

[−] Abstract and Keywords

Montague proposes a computational model to understand addiction. His key idea is a special kind of reward prediction error signal in addicts. In his comments, Yaffe discusses what Montague's work on the neuroscience of addiction does and does not show about moral responsibility. Sripada then outlines how additional deficits in reflective judgments of addicts might also be relevant to their moral responsibility. Montague replies by agreeing that we need a new generation of models to capture the kinds of considerations raised by his commentators.

Keywords: Free will, Moral responsibility, Addiction, Reward prediction error, Computational models

> Those are my principles, and if you don't like them, well, I have others.
>
> —Groucho Marx

Like Groucho Marx, most humans have opinions about their choices—strong opinions in many cases. This claim can be tested by simply asking anyone you encounter why they made some life choice (choice of mate, choice of job, place to live, etc.), and invariably there will be a response. The character and content of these opinions will likely vary widely, but there will be opinions. A similar experiment asking people about the role of personal decision making in quantum mechanics or the mechanistic role of the neurotransmitter dopamine in choice tends to yield silence. The origin of this difference seems obvious—we live intimately within our own behavior and thoughts day-to-day whether or not we have the interest or training to have developed a rich narrative about the role of choice in quantum mechanical experiments. The difference here is stark and represents a kind of narrative gap, which I believe is a barrier to our efforts to make third-party scientific accounts of what could be called willful choice. The shallow goal for this

chapter is to sketch briefly the way that neuroscience experiments frame human decision making and to use this to illustrate the ways that the human ability to exert control over their actions fails in addiction. This will allow us to point at many features of addiction as changes in the ability to exert cognitive control over choices.

A Classical View of Decision Making and Its Connection to Learning

In the usually practical world of neuroscience experiments, detailed philosophical accounts of free will do not tend to be part of experimental design. Only time will tell as to the utility of this omission. Instead, neuroscience experiments decompose decision making in humans in a classical way **(p.280)** where a decision-making agent (1) frames a problem (picks a representation), (2) values the states available to it, and (3) maps the states and valuations to some action or change of state. Before unpacking the details of these steps a bit more and relating them to computational models, let us emphasize that this depiction of choice is inherently sequential with each step following on the heels of the next. From a computational perspective, this implies some degree of statistical independence with time acting as the independent variable—we will return to this notion below. Also, the decision maker in such a scheme is usually rendered as a rational agent that follows some kind of maxim—like maximize one's expected utility over the available choices. This setup for decision making would be recognizable to anyone schooled in the basic canon for rational choice theory from the mid-twentieth century (e.g., von Neuman & Morgenstern, 1947; Savage, 1954; Simon, 1956; Luce & Raiffa, 1957). It is this latter rendering that has allowed for deep connections to computational models from the optimal control, reinforcement learning, and experimental psychology literatures (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). These literatures can now be related to some identifiable neural substrates, a fact that should enliven and expand any discussion of the concept of free will.

We will set up the decision-making setting, outline the computational models used in this context, and then ask how these models inform the way that choice—or more loosely changes in the freedom to choose—is thought to be perturbed in the addicted state. We will see that constraints on our current understanding about free will flow from several very different perspectives including known neurobiological substrates affecting choice.

The Markov Setting for a Learning Agent

The setting is an agent moving about an environment and learning about rewards in that environment and the sensory experiences that predict rewards. These agents may have internal models (stored models) of themselves and their environs, which they can update through experience. As outlined above, decision making is operationalized as being composed of three basic functions—(1) framing, (2) valuing, and (3) choosing. The first term means the activation, computation, or recall of some kind of state space for representing the problem at hand. The idea is that distinct frames act as representations distinguishable from one another. The second step involves valuation over the frame—typically depicted as a way to assign value to the states of the agent. The third step is the mapping from states **(p.281)** and values to actions—which themselves change the state of the agent. This is a typical breakdown of how neuroscience experiments on motivated choice conceptualize the problem. Neurally, very little is known about the framing part in detail. Valuation is the step where there is now a growing body of data that connects identifiable neural systems (dopamine systems) to the computation of values

associated either with sensory cues or actions (Montague, Hyman, & Cohen, 2004; Daw & Doya, 2006). Let's recapitulate with a more concrete example.

Imagine a mobile creature moving around an arena transitioning from one state to another. For the purposes of this chapter the nature of these states does not matter. In reinforcement learning models, the value of a state is explicitly defined as the average reward expected from that state forward into the future (Montague et al., 2004; Daw & Doya, 2006; Dayan, 2012). In our arena, let's imagine that there are food pellets hidden under the bedding covering the floor and that they are hidden in consistent locations from one learning trial to the next. The agent (here thought of as a rat) moves around, absorbs sensory information, and every now and then encounters a pellet. We know that the rats will learn a model of where to expect the food pellets in future trials as long as the pellet positions are signaled by a consistent set of sensory experiences. Suppose we now replace one of the food pellets with a pellet soaked in cocaine and keep this pellet in the same location from trial to trial. The animal will continually revisit that location with a dramatically higher frequency than the other locations—it becomes a highly valued place to that animal. Even if the cocaine pellet is removed entirely, the addicted animal will revisit that place in the arena with great eagerness and may even do so while ignoring the other pellets. Eventually, this behavior may diminish or even halt altogether depending on the behavioral arrangement, but the cues associated with that location in the arena have become highly valued to the animal even when there is no obvious return to the animal to warrant such valuation. This is one of the problems of addiction—the overvaluation of the cues surrounding the drug-taking experience.

What do we know about the neural processes that help support such a behavioral cul-de-sac? We actually know a great deal at many scales including how addicted behaviors reinstate once drug taking ceases or is extinguished. This is a gigantic literature too large to skim here. For our purposes, we are left with a question: Once the cocaine pellet is removed, we are left to wonder whether the animal can exert control over its behavior since revisiting the cocaine-pellet site yields nothing for the effort. Has it lost control? Not entirely. One can show that the animal can control its **(p.282)** behavior in other tasks. Has it lost control when pitting a visit to the former place of the cocaine pellet against other locations that are likely to contain a pellet? It depends on how one ranks the state of the animal. In the case that the animal is hungry, then visits to the overvalued location that do not yield a pellet are irrational but perhaps not so costly as to declare a complete loss of control.

One point to make for this example is that the idea of free control—let's call it the capacity to choose to ambulate to any location in the arena (after learning the cocaine-laden pellet location)—has been diminished in part because of the ecologically nonsensical valuation function that has developed for the rat. It is well-known that cocaine slows the reuptake of the neuromodulator dopamine and thus potentiates its signaling. Dopamine is critically involved in the pursuit of appetitive rewards although it is not the only neurochemical player in this domain. Over the last 20 years a computational account of the information encoded by dopaminergic activity in neurons in the midbrain shed even more specific insights into these behavioral facts illustrated in our example.

The Overvaluation Model and Its Implications for Diminished Control

In the simple example above, the way that control is diminished is through an assignment of excessive value to the cues associated with drug taking. This is a feature of addiction common to all drugs of abuse or even behaviors considered addictive. The model of addiction presented below derives from a now common way to frame the output of dopamine systems from a computational perspective. Dopamine is a potent neuromodulator involved in a number of important cognitive functions. In humans, dopamine is produced and released throughout the nervous system by two small collections of neurons in the midbrain called the substantia nigra and ventral tegmental area. Other than a small projection from the hypo-thalamus to the pituitary gland, these midbrain dopamine neurons are the only source of dopamine delivery throughout the brain. The importance of dopamine as a neurotransmitter is highlighted by the disorders that involve dopamine—drug addiction, Parkinson's disease, and various forms of psychosis. Here we focus on the role of dopamine in drug addiction because it's here that prevailing computational models of dopamine function shine some light on the issue.

One computational hypothesis is that dopaminergic systems encode a reward prediction error signal in modulations of their spike output—a model that matches a large amount of empirical data **(p.283)** (Montague et al., 1994, 1995, 1996, 2004; Daw & Doya, 2006; Dayan, 2012a, 2012b; Dayan & Walton, 2012; also see Bayer & Glimcher, 2005). And while this account certainly does not encompass all the functions played by midbrain dopamine signaling, it provides a very granular view of how specific variables influence valuation in the area of appetitive learning. It also points the way to a very specific computational model for addiction—one that exposes many subtleties in the question of willful choice. The error signal alluded to above derives from an explicit goal of learning in the model—that is, the goal is to use information from the environment (garnered typically by exploration) to learn a value function over states. The value of each state is taken as the expected value (average value) of the discounted reward from that state into the distant future as described here:

$V(s) = E\{r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots\}$ when $s_t = s$ and $0 < \gamma \leq 1$. Here V (S) is the value of state *s* and *E* is the expected value of the included expression. The natural prediction error emerges from this formula by noticing that the value of states at successive times obeys a recursion relation: $r_t + \gamma V(S_{t+1}) = V(S_t)$. Thus the difference between these expressions represents any discrepancy in the valuation of states but acts as a reward prediction error signal:

$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t)$. This error signal can be used to update predictions about the values of states and even to inform and improve the mapping from states and values on to actions. This has been a very successful framework in providing a principled connection between an important biological system that influences learning and decision making and computational principles with guiding equations that can be used to explore consequences in novel settings (Montague et al., 1994, 1996, 2004; Daw & Doya, 2006; more recently, Dayan, 2012a, 2012b). The model does not account for all-things-dopamine, and many interesting phenomena that depend on dopa-mine escape some of its predictions (Berridge, 2007). Nevertheless, the basic model provides a very detailed way to understand the overvaluation feature of drug addiction—as mediated by dopamine signaling. And over-valuation would seem like a dramatic change in willful choice, all other features of the creature's repertoire remaining intact. Let's see how this is possible.

In 2004, David Redish used the above model to propose one specific way that the computational account of dopamine signaling would mediate **(p.284)** the overvaluation of drug-related cues and thus provide a conduit for the aberrant behaviors that follow (Redish, 2004). As before, the value of a state is the integrated discounted rewards from that state forward but expressing this as a continuous process: $V(t) = \int_t^{\infty} d\tau \cdot \gamma^{\tau - t} E[R(\tau)].$ γ is a discount parameter between 0 and 1. As an agent transitions from state *Si* to state *Sj* at time *t* and receives (observes) reward $R(S_j)$, the prediction error signal is $\delta(t) = \gamma^{delay}\left[R(S_j) + V(S_j)\right] - V(S_i)$ where the delay is the time spent in state S_i. The value of the state visited at time *t* is updated by an amount proportional to δ*t*. The main idea is that taking exogenous dopamine-enhancing substances like cocaine produces a noncompensable (Redish's term) dopamine increase that cannot be learned by the system, but instead an increase in value occurs without end. This scheme accounts nicely for why drug-related cues and even contexts are overvalued. The influence on actions that lead to these states is direct—the system will "overselect actions" (Redish, 2004; see Dayan, 2012a, 2012b) that lead to drug receipt and the subsequent uncompensated dopamine increase.

This model of addiction provides a detailed and parametric model for why drug cues are overvalued and why actions leading to drug taking would be repeated with ease. The model, of course, does not include all the complicated aspects of cognitive control that impinge on drug-acquiring actions; however, it does account for two important features of addiction in computational terms that enjoy their own guiding equations. So does Redish's model, based on the formerly proposed reinforcement learning model, account for changes in will or control? Yes. The model provides a detailed account in terms of discount parameters, state space transitions, and error terms that drive learning. In this sense, a psychologically infused discussion about decreases in control can be made much more precise, and the often blunter discussion about free will can be parameterized. These parameters now have the chance to be mapped back onto the measured function of midbrain dopaminergic systems. This is a very small step and possibly disappointing in its modest gains—however, it has the advantage of translating some aspects of cognitive control to biological substrates as modeled by principled computational models. These steps open up the possibility of understanding the problem of addiction more fully.

(p.285) We have focused here on the valuation part of our operational decomposition for human decision making. A computational model was forwarded that has deep connections to the optimal control literature and to extant data derived from many experiments on the dopamine system of mammals. These dual connections provided a new way to see the changes in control that result in addiction as a multipart problem with detailed equations that can be used to model changes in control in other novel settings. This discussion raises some of the more difficult issues surrounding cognitive control, and again dopamine plays a role, but in more varied contexts. A seminal paper by O' Reilly and colleagues first proposed analogous computational models but applied to the area of cognitive control (O' Reilly et al., 1999). In these models, many other variables are introduced that through interacting networks provide for a new understanding of control and loss of control.

This chapter has emphasized how an age-old question about free will and the possibility that it was “diminished” in addiction leads to very granular models where the question is no longer so singular. The chapter has also avoided commenting on the technical philosophical aspects of the question of free will since the models and our understanding of their relation to biology are in their very early stages. **(p.286)**

(p.287) 7.1 Dopamine Dysfunction and Addict Responsibility: A Comment on Read Montague's “The Freedom to Choose and Drug Addiction”

The path-breaking work of Peter Dayan and Read Montague, and others, on the role of the dopamine signal in valuation and decision making is of the first importance. Of particular importance is the work that has been done in properly characterizing dopamine's role *in functional terms*, rather than in purely biochemical terms.[1] Until descriptions of biochemical mechanisms are translated into the language of psychological or computational functioning, next to nothing can be said about their relevance to the nature or extent of freedom of will, or responsibility, or any of the other related notions that philosophers and others concerned with legal and ethical evaluation hope to understand. The bridge built by Dayan and Montague and others between neuroscience and computational models of learning and decision making has made such functional description possible. We are now in a better position to ask what our growing knowledge of the dopamine system tells us about freedom of will.

If Dayan and Montague are right, and there is a substantial amount of evidence favoring their view (much of which is cited in Montague's chapter), in healthy people the dopamine signal carries information about errors in valuation, about ways in which the rewards we have received as a result of our conduct differ from those we expected to receive and which we were motivated to achieve. This is unlikely to be the only function of the dopamine signal, but it appears to be one of them. If this is right, then the dopamine signal is important to decision making in large part because it is important to *learning*. Its import is not exhausted by its role, if it plays one, in moving an agent toward a goal on one, single occasion; it plays a crucial role in structuring motivation the next time the agent acts, having updated his conception of what values the alternatives promise in light of the outcomes experienced in prior action.

By describing the dopamine signal in terms of the information relevant to practical reasoning that is carried by it, Dayan and Montague's work **(p.288)** supplies a substantial constraint on our answers to a variety of questions of central importance to the assessment of responsibility for wrongdoing. This is particularly true of questions about the relevance of disorders of dopamine function, such as addiction, to responsibility. For instance: Are addicts in control when they choose to take drugs, or pursue them, often harming others in the process? Do they exhibit the kinds of faults that make blame and sometimes criminal punishment appropriate? Our answers to these questions must be at least consistent with the fact that addicts' brains do not function as they ought when it comes to the way in which they carry information about evaluative error. And there is a significant possibility that the way in which addicts process information about evaluative errors, or the way their brains represent such information, is more than just *consistent* with the right answers to such normatively important questions; perhaps it even provides a tool from which we can draw *insight* into how they ought to be answered.

It is tempting to hold that any condition (e.g., psychological disorder, immaturity, temptation, craving) that potentially ameliorates a person's responsibility for wrongdoing does so, when it does, thanks to the fact that it diminishes control over conduct. In fact, some of Montague's remarks in his chapter suggest that he holds such a position.[2] On such a view, the conceptual route from abnormality to diminished responsibility runs, inevitably, first through diminished control. On such a view, those who wish to assert that, for instance, an individual with obsessive–compulsive disorder has diminished responsibility for neglecting a child in order to wash his or her hands twelve times an hour must also assert that the obsessive–compulsive disorder diminishes control over conduct. To give this kind of pride of place to control is to hold that ultimately the question of how a condition like obsessive–compulsive disorder affects, or fails to affect, responsibility reduces to the question of how it affects control.

However, a small amount of reflection on the diversity of our judgments of the responsibility of others serves to identify at least two alternative, control-independent ways in which a person's condition can result in his or her being less responsible for wrongful conduct than he or she would have been in its absence. First, sometimes, thanks to the agent's special condition, refraining from wrongful conduct would require the agent to bear substantial burdens that those who are not in that condition need not bear in order to act as they ought. Victims of coercion, for instance, find themselves in such conditions. Those who are not under threat can typically avoid harming others without thereby, for instance, having their fingers broken; not so for those in the clutches of someone willing to do **(p.289)** that who has attached that nasty result to noncompliance with the demand for wrongful conduct on the agent's part. If the burdens are big enough, and the imposition of them is not the agent's fault, we take the agent to be diminished in responsibility for wrongdoing. Sometimes, in fact, we wipe the slate clean in such cases, holding the agent to be not just diminished in responsibility but entirely nonresponsible.

However, conditions that attach burdens to norm compliance that others do not bear do not in any literal sense limit the agent's control over what he or she does. The agent remains capable of suffering the burden instead of engaging in wrongdoing. True, there is something that people in such predicaments *cannot* do that others can: They cannot comply with the norm while at the same time avoiding the burden; they cannot do as they ought while at the same time avoiding injury to their hands, in our example. However, what makes this fact significant to our assessment of responsibility cannot be that it is a diminishment in *control*. After all, someone who will inevitably *benefit* from wrongdoing also has diminished control relative to someone who will not benefit; such a person cannot engage in wrongdoing and at the same time avoid benefits. But this fact does not show the person to be diminished in responsibility. The reason is obvious: What is significant about the unavoidability of packages of compliance or noncompliance with norms, on the one hand, and burdens and benefits, on the other, are the *burdens and benefits*, not the *unavoidability* of the packages. Moral and legal norms require us to, for instance, refrain from robbing banks. Inevitably, and in every case, there are unavoidable consequences, both good and bad, for compliance or noncompliance with such norms. But, still, those who can comply, can comply. If we take them to have diminished responsibility due to the burdens they must suffer in order to comply, it is because of the burdens, not because of the rather attenuated sense in which their control is diminished. Even those who *can* do as they

ought sometimes *cannot be expected to*. This idea is reflected in criminal codes, especially in affirmative defenses, such as the defense of duress.[3]

Second, there appear to be capacities that a person must have in order to be justifiably blamed and punished for wrongdoing, the absence of which are compatible with full control over conduct. The famous M' Naghten test for legal insanity is driven by this idea. A person could lack the capacity "to appreciate the wrongfulness of his conduct"[4] while still being entirely under control while engaging in it. Someone, for instance, in the grip of paranoid delusions might believe himself fully justified in killing someone whom he takes to threaten his livelihood. And, in such a **(p.290)** case, the delusions may be so pressing as to ensure that the agent will have the false belief that the killing of another is justified. He may entirely lack the capacity to recognize the wrongfulness of his conduct. And yet he might be perfectly in control of himself when he acts, or at least no less in control of his conduct than anyone else who kills another in order to protect his or her own interests and falsely believes himself or herself to be in the right. There is something that someone insane by M' Naghten's standards *cannot* do: In this example, he cannot recognize that he should not kill. But then this is true often enough of those who are not insane by M' Naghten's standards, also. Things strike us as they strike us, and most of us, often enough anyway, have little ability to see things otherwise, especially when emotions are high and time for reflection is short, as in many circumstances that give rise to crime. The capacity that the M' Naghten rules associate with sanity seems to matter to responsibility despite the fact that it might not be necessary for full control over conduct.[5]

So, when asking what the relevance is to freedom and responsibility of the dopamine signal's functional role, and its dysfunction in addicts and others suffering from disorders of dopamine signaling, we need to be alive to the possibility that it matters without bearing on control. Perhaps, thanks to dopamine dysfunction, addicts bear burdens that nonaddicts do not bear. Perhaps they have what we might call a "burden-based excuse." Or perhaps, thanks to dopamine dysfunction, they lack a basic capacity for subjection to blame or punishment of a sort that is required for responsibility despite not underlying any form of control over conduct. Perhaps, that is, they have what we might call a "normative incompetence" excuse.

It is important to see that if Dayan and Montague are right, and the dopamine signal plays a crucial role in evaluative learning that is undermined in addiction, it remains nonobvious what bearing addiction has on responsibility. We do know some things. We know, for instance, that a deficit in a capacity to learn the values of alternative outcomes does not weaken control over conduct, at least not under many appealing accounts of the nature of control. Consider, for instance, the popular account of control in terms of reasons-responsiveness.[6] Under such a view, roughly, a person's conduct was under control just in case he or she engaged in it for reasons and would have acted differently had reasons favored an alternative. The fact that a person values cocaine consumption, and continues to value it, even increases his or her valuation of it, following consumption and despite the mismatch between the goods that he or she takes it to promise and those that it actually provides, does not imply that the person fails to consume it for reasons, or that he or she would still consume it **(p.291)** even if there were reasons not to. Even given their deficits in evaluative learning, addicts do not assign such high value to drug consumption that they would continue to consume *no matter what*. Provide them with powerful reasons to refrain from consumption, and they refrain. They take drugs for reasons and are

responsive to countervailing reasons, even if less responsive than those who are better at learning to value actions and outcomes. They are responsive to reasons in the same way as anyone who cares a lot about something. (Compare to someone who loves golf.) Thus, they are in control of their conduct by the standards of at least one popular account of the nature of control.

It is true that addicts respond to reasons in a distinctive pattern, a pattern that differs from that of nonaddicts. There are things that they will do in order to consume drugs that nonaddicts will not do, or will need different, or greater, incentives to do. And, in general, they require much stronger reasons to prompt nonconsumption than nonaddict, recreational drug users would require. Perhaps these differences in the ways in which they respond to reasons shows that addicts have less control than nonaddicts. We could, of course, craft a graded conception of control under which one has a lesser ability to do something if one would need greater reasons to do it before one would recognize or respond to such reasons. However, it is far from clear that degrees of responsibility correspond at all with degrees of control in this sense. Someone who needs greater reasons to refrain from hurting someone else might, in a sense, have diminished control relative to those who will refrain even in the absence of such strong reasons. But that hardly implies that the person is less responsible for harming another when he or she does.[7]

No, if addicts' deficits in evaluative learning, rooted in dopamine signal dysfunction, matter to responsibility or freedom of will, it must because they matter for reasons that are independent from control. Elsewhere I have argued that what these deficits show is that addicts bear burdens for compliance with norms that the nonaddicted do not bear.[8] In particular, unlike those of us who are not addicted, sometimes the only way for an addict to avoid wrongdoing is to sacrifice his or her autonomy, understood as the capacity to act in accordance with what one values most. Given that addicts are, inevitably, going to (even if temporarily[9]) value drug consumption over refraining from drug consumption, they can avoid consumption only by acting contrary to what they most value at the time of action. Sometimes this requires addicts to cede control of their behavior to others. The alcoholic who has to pass the corner bar if he is to meet his obligations to his friends or his children may need to let someone else drive him past **(p.292)** the bar. If he maintains control over his movements, he will inevitably stop at the bar and shirk his obligations. He *can* meet his obligations because he can hand himself over to someone else who will see to it that he meets his obligations. His abilities are not undermined; he does not have diminished control. But he does bear extra burdens thanks to his addiction. Most of us need not give up our autonomy in order to meet our obligations. And giving up one's autonomy is a burden, a substantial burden, in fact, that liberal societies do their best to protect people from having to bear. Thus, there is reason to think that the functional role of the dopamine signal, uncovered by the work of Montague and others, suggests that addicts have a burden-based excuse for their bad behavior.

The line of thought just sketched is not at all decisive. There remains the possibility, for instance, that addicts have normative incompetence excuses either instead of, or in addition to, their burden-based excuses. To know whether or not this is true, we would need to know more about what, exactly, is involved in normative competence. The M' Naghten formula is vague and imprecise, and there is much more detailed work on the issue to be found, and to be done.[10] In

addition, as Montague emphasizes, and as has been alluded to above, to say that the dopamine signal plays a role in evaluative learning that is perturbed in addiction is not to say that it does not play other crucial roles that are also perturbed in addiction. Perhaps we will find that the other roles that it plays are essential to control, either in the sense of reasons-responsiveness or in some other sense. Or perhaps we will find that it plays other roles that are determinative of other burdens that we must bear, or can avoid, through action. Or perhaps it will bear on normative competence in ways that we are not yet in a position to appreciate. There is a great deal more work to be done, both on the neuroscientific side and on the philosophical side. We need not just more information about the biochemical facts, nor just more devices for translating those facts into the language of computational and psychological functioning, but also more discoveries about the precise features that are relevant to moral and legal evaluation.

Notes

(p.294)

(p.295) 7.2 The Second Hit in Addiction

To make sense of drug addiction, we must postulate at least two pathologies. The first is in the realm of brute desire. Humans want various things. We want food and sex, as well as such things as chocolate and video games. In addiction, the desire for drugs is excessive in strength and is likely excessive other ways as well (e.g., the desire is excessively persistent and pervasive). Drug seeking and drug consumption crowd out the pursuits that we consider central to a well-led life—things such as love, family, work, and material well-being.

However, addiction must also involve a second abnormality. Humans have the ability to exercise restraint with respect to their desires, including their extremely strong desires. Even when I am absolutely famished, I can stop myself from eating; when concupiscent, I manage to restrain myself from leering or worse. Put another way, humans not only desire to do this or that, they form reflective judgments about what it would be desirable to do, and such judgments have the ability to reign in their appetites and steer them clear of temptation.[1]

The conflict between reflective judgment and appetites is a central feature of addiction. Addiction does not simply involve wanton, unregulated pursuit of drugs. Addicts articulate a desire to quit—indeed they vow in the most forceful language that this high will be their very last. They castigate themselves for caving in. They attend time-consuming therapy sessions and self-help meetings. They enroll in residential programs that extend for months and cost tens of thousands of dollars. In short, they do just the things one would expect them to do if they *sincerely* judge that they should quit and are committed to doing so. Yet somehow, their judgments about what it would be desirable to do fail to govern their actions. Thus in addition to excessive desire, addiction appears to involve a second pathology—the failure of addicts' capacities for reflective judgment to regulate their wayward desires. This is what we might call the “second hit” in addiction.

(p.296) The contest between reflective judgment and wayward desire is a key issue in understanding moral responsibility in addiction. Merely having desires, even very strong desires, does not erase moral responsibility for actions. Mother Teresa had strong desires to help the suffering, and surely she was morally responsible for her charitable actions. My aim is

to explore how reflective regulation of strong desires, and in particular its failure, might illuminate the question of moral responsibility in addiction.

Montague's Research on Dopamine Signal Dysfunction in Addiction

Over the last decade, Read Montague and his collaborators have conducted seminal research investigating the neurobiological basis of addiction. They have developed sophisticated computational models of reward learning drawing on theoretical work in artificial intelligence and computer science. A central player in these models is the neurotransmitter dopamine, which is hypothesized to encode a prediction error signal. In particular, phasic emission of dopamine informs learning systems that the current state is "better than expected," leading to an increase in the value attached to this state. This signal is critical to ensuring that valuations of a state are closely tied to the actual receipt of future rewards. Drugs of abuse are potent brain releasers of dopamine, and in exogenously manipulating the dopamine signal, they produce a perverse cascade of consequences. With each episode of drug consumption, dopamine is released, the current state (the state of consuming drugs) is regarded as better than expected, and hence the value attached to this state is commensurately increased—even in the absence of any real downstream rewards. With repeated drug use, the result is profound hypervaluation of drug consumption.

Where does Montague's neurocomputational account of dopamine dysfunction fit within the two-factor model of addiction proposed above, that is, the model that distinguishes between excessive appetitive desires and inadequate control over these desires by one's reflective judgments? It seems that Montague's neurocomputational account is most directly related to the first factor. His account offers a satisfying and computationally precise account of why addicts experience excessive motivation directed at drug consumption. It specifically helps to make sense of one aspect of the irrationality of addiction—the fact that addicts' valuations of drug consumption are out of step with the rewards that consuming drugs actually produces. This is explained in terms of the exogenous manipulation of dopamine that "tricks" the brain into assigning enormous value to a state, that is, consuming drugs, that yields few actual rewards.

(p.297) What then of the second hit in addition? I have noted that humans have the ability to exercise restraint with respect to their desires, including their extremely strong desires. An account is still required of how regulation, and in particular its failure, plays a role in sustaining addiction. In what follows, I will sketch four models of regulatory failure in addiction. In proposing these models, I am guided by Montague's penetrating research into the appetitive dysfunction in addiction. I will also draw on Gideon Yaffe's insightful comment on Montague's chapter, and in particular his distinction between *burden-based* excuse and excuse due to *normative incompetence*. It is hoped that as we explore these four models, additional insight will be gained about how addiction might mitigate or in some cases eliminate moral responsibility.

Four Accounts of Regulatory Failure in Addiction

The first account of failure of regulation in addiction is the *irresistible impulses* model. According to this model, dopamine signal dysfunction in addiction produces drug-directed desires that are so hypertrophied, the addict literarily cannot resist these desires. The irresistible impulse model has been criticized in detail elsewhere (see, e.g., Husak, 1992; Morse, 2000), and I will not restate its problems at length; just a couple of brief observations will

suffice. First, addicts, even those suffering from extremely severe addiction, *do* have the ability to exercise restraint with respect to their drug-directed desires most of the time. They repeatedly quit and have some success (though often short-lived). The fact that they have any success at all suggests the irresistible impulses model is inadequate and a more nuanced picture is required. A second relevant observation draws on imaging data about synaptic dopamine release due to natural rewards such as food and video games. Endogenous dopamine release in the context of these natural rewards is often found to be comparable in magnitude to exogenous release due to drugs of abuse (see, e.g., Koepp et al., 1998). Given that desires directed at natural rewards are resistible at least *most* of the time if not *all* the time, it seems unwise to conclude on the basis of existing neurobiological evidence that desires directed at drugs of abuse are somehow very different and are not, at least most of the time, resistible.

Were the irresistible impulses model after all correct, it might seem that the implications for moral responsibility are relatively straightforward. Moral responsibility, one might think, requires *control* over one's desires and actions, and if addicts' desires are literally irresistible, then they lack **(p.298)** the requisite control. But we should be wary, I think, of concluding too quickly that control is the critical feature here that undermines moral responsibility. Harry Frankfurt's example of “willing” and “unwilling” addicts illustrates this point (Frankfurt, 2003). Consider two addicts, both of whom have an irresistible desire to use a narcotic. The unwilling addict rejects his addiction and desires that his desire to use the narcotic be extinguished. The willing addict endorses his addiction, and were his desire to use the narcotic ever extinguished, he would seek to reinstate it. The addicts do not differ in terms of control; for both addicts, the desire to use the drug is equally irresistible. However, there is a strong intuition that the addicts differ in terms of moral responsibility. The willing addict is morally responsible for using the drug while the unwilling addict is not, or at least the two addicts differ in their degree of moral responsibility. Elsewhere, I and others (Frankfurt, 2003; C. Sripada, 2013) have developed accounts of moral responsibility that try to explain why our responsibility judgments of the willing and unwilling addict differ. However, for now, I leave it that regulatory failure due to irresistible impulses can, at least in some cases, undermine moral responsibility. I leave it open however whether the pathway by which it undermines responsibility goes through the absence of control or through other factors.

The second model of regulatory failure in addiction says not that resistance is futile, but rather that it is *difficult*, perhaps excessively so.[2] It is a folk platitude that exercising willpower to control one's own desires is effortful and fatiguing. This picture is vindicated by recent psychological research that supports a “limited resource” model of regulatory control (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Hagger, Wood, Stiff, & Chatzisarantis, 2010). Engaging in various kinds of self-regulation—controlling one's urges, thoughts, or habits—depletes a limited mental resource, making it harder for the person to perform effectively in subsequent tasks. These findings suggest a model in which addicts fail at regulation not because they literally *can't* regulate, but because at some point, it is *too hard*. Enduring battles with drug-directed urges leave addicts fatigued and depleted. Other tasks that require self-regulation—dealing with home finances, parenting a fussy child, putting up with an annoying boss—become significantly more challenging. As the number of activities that draw on one's self-regulatory capacities increases, at some point addicts find it too difficult to continue to exert regulatory control over their thoughts and their desires and instead give in.[3]

Might this *too-difficult-to-resist* model of regulation failure be used in addiction as a basis to mitigate, or even erase, moral responsibility? Yaffe **(p.299)** provides a defense of what he calls "burdens-based" excuses. For example, a person who is coerced for the purposes of getting him to ϕ rather than ψ might bear a cost of broken fingers should he choose to ψ. A person who is not the victim of coercion can choose to ψ without bearing this burden. Examples of coercion remind us that when the burdens are sufficiently high, we take the agent's moral responsibility to be mitigated.

With regard specifically to addiction, Yaffe draws attention to certain *behavioral restrictions* imposed on the addict, if he is to maintain sobriety, that diminish his autonomy. He offers the example of the alcoholic who cedes driving duties over to a friend lest he be tempted to head to a favorite bar. The model of regulatory failure in addiction presently under discussion—the too-difficult-to-resist model—suggests that there may be additional *psychic* costs to maintaining sobriety, on top of the autonomy-threatening behavioral restrictions that Yaffe highlights. Addicts have to expend tremendous effort battling drug-directed urges, leaving them drained and making the performance of other day-to-day tasks that require self-regulation more challenging. As the cumulative tally of these burdens gets sufficiently high, the moral responsibility of the addict for failing to maintain sobriety might well be mitigated.

Let us turn now to a third *epistemic limitations* model. This model says that addicts have the capacity to regulate their drug-directed desires, but they fail to exercise this capacity when they need to because of cognitive limitations or errors, such as biased beliefs, distorted predictions about the future, or faulty normative judgments. The epistemic limitations model encompasses a wide variety of factors, so let me briefly develop one specific version of this sort of account.

We know that addicts engage in steeper temporal discounting of the future than nonaddicts (Bickel & Marsch, 2001; see also Yaffe, 2001). One way to understand this phenomenon is as a form of *myopia*—it is not that addicts don't *care* about consequences in the distant future but rather that they can't adequately mentally represent those distant consequences so that they could even be cared about; they are in a sense *future blind*. Evidence for this view comes from Luhmann and colleagues (2008), in which participants were presented with choices between smaller earlier rewards and larger later rewards during functional imaging. They found that individuals who scored highest on a measure of impulsivity exhibited significantly reduced activation in brain regions in medial frontal and lateral parietal cortex associated with prospection, internal mental simulation of future states of affairs. Though this finding admits of several interpretations, one possibility suggested by the authors is that steeper temporal **(p. 300)** discounting in impulsive individuals arises due to diminished functioning in brain regions that allow a person to mentally represent the future.

Yaffe suggests that morally responsible agency requires that a person be *normatively competent*. He notes it is difficult to say comprehensively what all normative competence amounts to. However, in addition to having the appropriate moral knowledge, that is, knowledge of right and wrong, it is plausible that a person must also possess various kinds of nonmoral, purely descriptive kinds of information. For example, Aristotle argued that to be morally responsible for an action, individuals at least must know what they are doing, what or whom they are acting on, and the ends for which the action is done (Aristotle, 1999, 1109b30–1111b5).

If addicts fail to regulate drug-directed desires because of certain cognitive limitations that, for example, prevent adequate grasp of future consequences, then these knowledge conditions may not be satisfied, thus mitigating or removing moral responsibility.

The final account of regulatory failure in addiction differs markedly from the first three in that it does not propose that addicts' desires are somehow too strong or that their regulatory powers are too weak or otherwise impaired.

Let me begin with a bit of warning. In setting up this account, I will be presenting a highly simplified picture of addiction. My aim is not to ignore, obscure, or otherwise trivialize aspects of this complex and multifaceted disorder. My aim rather is to focus on just one phenomenon in addiction that, unless simplified and specifically showcased, can too easily be missed. Once the simple model is presented and the phenomenon I aim to highlight is made clear, it should be possible to add back nuances and complexities while retaining what I take to be the model's fundamental insight.

Consider an agent who experiences an active desire to ϕ. Next, a two-stage process ensues. First, the agent engages in deliberation and forms a reflective judgment that ϕing is *not* the thing he ought to do. Second, he or she performs various regulation strategies. These strategies might include such things as redirecting attention away from things in the environment that remind him of ϕing, mentally distancing himself from the hedonic appeal of ϕing, directly inhibiting the motivation to ϕ, and so on. Call these two stages together—the forming of the reflective judgment that opposes ϕing and the subsequent execution of ϕdirected regulation strategies—"*JR*."

Undertaking *JR* is not a trivial endeavor. A suite of sophisticated component processes (attention, working memory, practical reasoning, **(p.301)** simulation/prospection, inhibition, and many others) need to be activated and orchestrated in a coordinated way in order for *JR* to be successful. Now any complex and articulated process such as *JR* will inevitably have some rate of failure. The science of operations research tries to engineer industrial processes that minimize mistakes and mishaps. However, there is no known method to drive the rate of failure to zero. Thus it would be miraculous indeed if evolution fashioned from frail neural flesh sophisticated and articulated processes that exhibit no tendency to fail whatsoever.

So *JR* has some nonzero rate of failure. The next step is to pick a number to represent this failure rate. However, before we do this, let us be perfectly clear about what this number represents. This is not the rate at which we succumb to weakness of will[4] or have a change of heart and decide to indulge our temptations. Rather, this is the rate at which a person who *sincerely* judges that he or she ought not follow some wayward desire, and who has *no deficits* at all in his or her regulatory capacities, and who *fully* and *whole-heartedly* attempts to suppress the wayward desire, will instead end up failing due to intrinsic limitations in the reliability of the *JR* process itself. To mark this special circumstance, I will call these events "pure intrinsic failure events." Given the effectiveness with which we regulate desires in our day-to-day life, let us assume the rate of these events is quite low, say 0.001%. In other words, when we confront some temptation-directed desire that we sincerely oppose and fully resist, there is a one thousandth of a percent chance that the reflective processes by which we regulate the desire,

that is, *JR*, will exhibit a pure intrinsic failure and, for this reason, the wayward desire will win out.

The final step is to select a frequency with which our hypothetical agent will confront what I shall call "*JR*-eliciting desires." These are defined as desires with the following property: Were the agent's judgment and regulation processes functioning *flawlessly*, the agent would form a judgment that opposes the desire and would execute regulation strategies that suppress the desire. Most of our desires, of course, are not *JR*-eliciting desires. These are desires we reflectively endorse (e.g., the desire to drink fluids when thirsty), or at least they are desires that we reflectively *choose* to indulge. In assessing the intrinsic failure rate of judgment and regulation systems, it is *JR*-eliciting desires specifically that are the relevant category of interest.

To fix intuitions, let us start with a case of "regular," desires in the nonaddict. Suppose some man periodically craves salty deep-fried potato chips. Every other day, he has *JR*-eliciting desires to go to the vending machine at work and get a salty snack. If we assume the failure rate **(p. 302)** specified above (0.001%), then the man will suffer about one pure intrinsic failure event every five years.

But what happens if we increase the frequency with which a person must confront *JR*-eliciting desires? For example, let us suppose an addict experiences drug-directed desires six times a day (note: we are assuming that *all* of an addict's drug-directed desires are *JR*-eliciting). Then, according to our model, remarkably, the addict will suffer an intrinsic failure event every four to five months. Put another way, no matter how well-intentioned, and no matter how otherwise well-functioning his judgment and regulation processes, the addict is set up to relapse roughly twice a year, and indeed we should be surprised if he doesn't.

Is this assumption about the frequency of the addict's drug-directed desires at all reasonable? That is, do episodes of drug-directed desires occur in addicts with much greater frequency than corresponding *JR*-eliciting desires for more mundane things in healthy individuals, such as the desire for salty snacks? I believe the answer is yes. While I will not provide an exhaustive review, I will briefly sketch at least one line of evidence that favors this view.

Clinicians have long noted phenomenological similarities between addiction and obsessive–compulsive disorder (OCD). Like sufferers of OCD, addicts experience repetitive, intrusive thoughts and urges, and preoccupation with these thoughts and urges can, at least in some cases, occupy much of their day. In order to measure this phenomenon of *obsessionality* in addiction, alcohol researchers have adapted the Yale-Brown Obsessive-Compulsive Scale, the standard scale for OCD, to quantify repetitive drug-directed thoughts and desires. The resulting Yale-Brown Obsessive-Compulsive Scale for heavy drinking (YBOCS-hd; Anton, 2000) is now the most widely used scale to measure clinical phenomenology in alcohol use disorders and has been shown to reliably predict addiction severity and relapse (see, e.g., Schmidt, Helten, & Soyka, 2011). This suggests that obsessionality is indeed a fundamental aspect of addiction, and the supposition that addicts battle drug-directed desires multiple times a day is not at all implausible.

I dub this fourth model of regulatory failure in addiction the *cumulative failure* model. The model shows that when urges are frequent enough, even when the "*point" probability* of failure at any individual regulation event is tiny, the *cumulative probability* of failure rises over time to near certainty. I believe that it is fairly clear that the addict described by this fourth model is *not* **(p.303)** morally responsible for his relapsing into drug use. I will not, however, try to offer a comprehensive account of precisely *why* the addict fails to be morally responsible. In my view, standard accounts of moral responsibility, such as the popular reasons-responsiveness view[5] (Fischer & Ravizza, 1998), fail to provide an adequate explanation for why this fourth addict is not morally responsible. Elsewhere I propose an alternative account of moral responsibility that, I believe, does provide a satisfying explanation (C. Sripada, 2013).

A perhaps somewhat troubling feature of the cumulative failure model is that it explains breakdowns in regulatory control in addiction without postulating any of the familiar kinds of agentive deficits found in the first three models. The addict in the fourth model certainly does not suffer from *irresistible impulses* in any obvious sense. We have not proposed that his or her drug-directed desires are particularly strong, and in any case they are not proposed to be any stronger than the desires of the person who craves salty snacks. Moreover, the addict's success rate in resisting impulses is nearly perfect (99.999% successful)—the same as the nonaddicted individual. The same applies to the factors proposed in the next two models. That is, the addict in the fourth model has no greater *difficulty* in resisting any one of the *JR*-eliciting desires he or she confronts than a nonaddicted individual, and, it bears mentioning again, succeeds in resisting in nearly every single case. Additionally, the fourth model does not saddle the addict with any *epistemic deficits* such as distorted thinking or myopia for the future. In all these respects the addict in the fourth model is *exactly* like a healthy person without addiction. The only respect in which this fourth addict differs from a healthy individual is the frequency with which he or she battles episodes of *JR*-eliciting desires. This alone was sufficient to explain why, despite the addict's sincere and resolute judgment that he or she should not use drugs, the addict nonetheless relapses into drug use again and again and again.

Conclusion

Read Montague's pioneering work has shed light on abnormalities in addiction in the formation of drug-directed desires. However, addicts also exhibit a second pathology; their reflective judgments somehow fail to regulate their drug-directed motivations. This is the second hit in addiction. Four models have been formulated to account for failures of regulatory control in addiction. Three models propose agentive deficits that are standardly associated with addiction while the fourth does not propose any obvious agentive deficits at all. Each model suggests a different basis for mitigation or outright denial of the addict's moral responsibility.

Notes

(p.305) 7.3 Responses to Yaffe and Sripada

Sripada and Yaffe highlight a number of important features of drug addiction where our ignorance of the underlying computational issues and their neurobiological underpinnings hamstrings our ability to understand exactly how to think about desire and cognitive control. It

goes without saying that this same ignorance makes difficult the mapping from neural competency to responsibility. Two themes pertinent to our state of knowledge in neuroscience emerge from their commentaries.

Response to Yaffe

The first, offered in clear terms from Yaffe, is that issues of behavioral control can be seen as irrelevant for the purpose of understanding neural competencies that could (in principle) underwrite norm compliance. He focuses on the fact that all observable actions are associated with different benefits and burdens. Persons bear different (internal) burdens by virtue of the way their nervous system values actions, ideas, contemplated actions, and so on. Obviously such burdens could be changed by disease or injury. In this context, Yaffe leaves the neural difference between a burden and a benefit as a theoretical claim and focuses instead on the issue of whether such differences, should they exist in a meaningful sense, have any implications for how we rate a person's capacity for control. There are subtleties lurking here.

Yaffe supports his view by describing a clear mapping from the (internal) benefits and burdens to (observable) actions—that is, for any set of observable actions he claims that individuals can bear different internal burdens. For brevity, I am letting “burdens” substitute for what is clearly a signed quantity (benefits and burdens). Yaffe goes on to point out that all observable actions have the property that they are unavoidably linked to different burdens for different subjects. This setup allows him to claim that our **(p.306)** assessment of the responsibility of a subject for a given action (that may break a law or deviate from some other social norm) cannot depend on some idea that the subject's control is diminished. To summarize—there are internal burdens that vary from subject to subject for any given action. Hence, the assignment of responsibility for the outcomes of actions (e.g., those that transgress with respect to the law) must focus on the nature of the burdens and the ways they can come to differ across subjects. Yaffe's framework is important for drug addiction because the idea of diminished capacity for an addict then has nothing to do with any metric defined on the space of their possible actions but instead relates strictly to some assessment of their internal burdens (presuming we could measure them in a manner relevant to the law).

With this in mind, could we use modern neuroimaging and/or some kind of computational model to develop an eavesdropping method for the internal burdens? For a drug addict, this might be a method to assess covertly their degree of craving in some situation. While this seems possible, it has not yet proven feasible primarily because measures of craving invariably rely on some kind of conscious report. The state-of-the-art neuroimaging experiments currently have no capacity to relate conscious perception of craving to other craving conditions that may impinge on actions but not be apparent in conscious reports. This is just one subtlety. However, Yaffe's framing of the internal burden issue raises more important scientific subtleties.

One crucial feature of the burdens argument is the relationship between internal burdens and external actions. From a scientific perspective there are two broad categories of claim here. The first is the implicit claim that a subject's actions are quite flexibly available in the face of a range of internal burdens.

In support of this rendering I quote Yaffe:

> However, conditions that attach burdens to norm compliance that others do not bear do not in any literal sense limit the agent's control over what he or she does. The agent remains capable of suffering the burden instead of engaging in wrongdoing. (p. 289)

Well maybe. Yaffe makes a clarifying claim, but it might not be reasonable once the connection between burdens and possible actions is made clearer by future work. Let's imagine that we have a mathematically defined space of burdens and a similarly defined space of actions. Without committing to other structures on these spaces, we can say that Yaffe's statement amounts to a *claim about the mapping from the burden space to the action* ***(p.307)*** *space*. Let's also stipulate that other capacities internal to subjects also map onto the action space, but we will ignore this possibility. I would like to consider an extreme case that tests Yaffe's ideas about where the assignment of responsibility should focus.

In case 1, let's imagine two subjects with identical burden spaces and further imagine that the natural variability in the mapping from burdens to actions is not identical across the two subjects. Suppose that for equivalent burdens across the subjects, subject 2 has fewer actions available as output and that *all* these actions cross some kind of legal or social norm boundary. If the assignment of responsibility depends on the burdens borne by subject 2, then we have a problem since subject 2's burdens are equivalent with those of subject 1. The problem could be said to arise from the mapping onto the action space, but certainly not the burden space. Ok, so just include the mapping from burdens to actions in an enlarged notion of burdens and continue to apply the Yaffe argument. This appears reasonable, but now we have extended the notion of how internal mental states enter into the problem of responsibility. It might also matter how the "faulty" mapping came to be established—was it due to expected and normal biological variability or was it the consequence of a lifetime of choices that froze some mappings in and excluded others? There are many subtleties here not yet within the scope of neuroscience, developmental biology, or computational science to answer. Therefore, Yaffe's claim provides a nice, but challenging starting point. It highlights the stark gap in our knowledge about healthy cognitive development and variability in the mature cognition that such development produces. What are the biological underpinnings of normative competence across any dimension?

Response to Sripada

Sripada introduces a second and important theme—the capacity to forecast the future and its role in governing the choices of the present. To display past and present ideas about regulatory failure in humans, Sripada describes four models of regulation failure—(1) irresistible impulses, (2) resistance is difficult (think Yaffe's burden argument), (3) epistemic limitations, and (4) cumulative stochastic failure (think death of a thousand small blows).

The irresistible impulses idea influences much of modern work on addiction. For drugs of abuse, the case is straightforward. A person trades his or her desire for an immediate drug experience for nearly anything else including possible gains in the future. One way to conceptualize this propensity is to say that the subject can no longer correctly or reasonably **(p.308)** value the near-term future, and there is now an entire industry built around what is called "the intertemporal discounting problem." Implicit in this framework is the (somewhat blunt) idea that addicts have a problem with affective forecasting ("What will the future feel like if I do *X* now?") and more generally counterfactual thinking ("If I do *X* now, then Y will happen later, and this is

bad”). These two ideas are hard to separate as typically conceived, but they highlight a problem that almost defines an addict, so it's difficult to view this as a model of addiction that could provide new ways to gain biological insight into the problem. Instead they are good frameworks for parameterizing the degree to which and dimensions along which someone has difficulty forecasting into the future in a way that intervenes meaningfully on present choices.

Sripada introduces the notion of an accumulative model, where failures to regulate across time and for whatever reason act through learning mechanisms to change the overall function of those neural systems impacted by drugs of addiction. This point of view is almost certainly closer to the truth, and it helps to reframe what he calls the “second hit” of addiction, where addicts appear to lose or suppress the counterfactual capacities alluded to above.

Overall, we do not have a good scientific account of human impulses, their normal variety, or the normal variability of their mapping onto possible actions. Sripada's discussion makes this quite apparent, and I have no clever answer on this account. His piece highlights the need for next-generation models that seek to capture the way that human nervous systems generate impulses and guide them through regulation. Like Yaffe, he has raised a deep question, since such impulses are impacted by the culture in which our nervous systems are embedded and we have no good account of how our biological limitations funnel and shape the influence of the surrounding culture. Thus the issue of free will and control still holds center stage for neuroscience and for its use and misuse in the realm of our everyday problems like drug addiction and the assignment of culpability.

Notes:

(1) . Cf. P. R. Montague, P. Dayan, and T. J. Sejnowski (1996), “A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning,” *Journal of Neuroscience*, 16, 1936–1947; W. Schultz, P. Dayan, and P. R. Montague (1997), “A Neural Substrate of Prediction and Reward,” *Science*, 275, 1593–1599; P. Dayan and M. E. Walton (2012), “A Step-by-Step Guide to Dopamine,” *Biological Psychiatry*, doi: 10.1016/j.biopsych.2012.03.008.

(1) . I am relying here on a *dual-systems* picture of motivational architecture that distinguishes between, roughly, a reflective system that implements practical reasoning and a reactive system that consists of such processes as emotions, drives, and cravings. Dual-system views are well supported in psychology (Metcalfe & Mischel, 1999; Hofmann, Friese, & Strack, 2009) and the neurosciences (Bechara, 2005; Montague, King-Casas, & Cohen, 2006; Sanfey & Chang, 2008). For more on the regulatory processes by which the reflective system inhibits the reactive system, see Gross (1998) and Phan and Sripada (2013). See C. S. Sripada (2010, 2012a) for discussions of motivational architecture and regulatory control from a philosophical perspective.

(2) . Montague is more explicit elsewhere that this is his position. See Read Montague, “Free Will,” *Current Biology*, 18, R584. The same position is taken for granted by Patricia Churchland and Christopher Suhler, “Agency and Control: The Subcortical Role in Good Decisions” (this volume).

(2) . Neil Levy offers an illuminating discussion of a model broadly similar to this one in Levy (2006) .

(3) . Cf. *Model Penal Code* § 2.09.

(3) . Depletion of self-regulatory capacities might *also* be taken to play an important role in the irresistible impulses model. Since the irresistible impulses model was already discussed, here I am focusing on the role played by depletion in making resistance *difficult*, though not *impossible*.

(4) . This language is used in the *Model Penal Code* § 4.01(1). Every jurisdiction in the United States uses a test of this sort for insanity although a minority allow, as does the *Model Penal Code*, for an alternative "volitional" conception of insanity under which a mental disorder can excuse by making it very difficult or impossible for its sufferer to act legally.

(4) . In philosophy, weakness of will is understood not as an instance of the kind of failure currently under discussion but rather consists in *freely* and *intentionally* acting on a desire that is in opposition to one's all things considered best judgment.

(5) . The literature on the insanity defense, and this conception of it as concerned with a kind of normative competence distinct from diminished control, is vast. For a start, see Gary Watson (2011), "The Insanity Defense" in *The Routledge Companion to the Philosophy of Law*, Andrei Marmor (Ed.), pp. 205–221, New York: Routledge; Stephen Morse (1994), "Culpability and Control," *University of Pennsylvania Law Review*, 142, 1587.

(5) . This model is discussed by Yaffe, though he does not specifically endorse this model.

(6) . For the canonical statement of this conception of control, see John Fischer and Mark Ravizza (1998), *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

(7) . Fischer and Ravizza hold that "moderate" reasons-responsiveness is necessary for responsibility, where the agent is defined as possessing such responsiveness when he or she would recognize and respond to some reasons to act, although not necessarily all, and in an intelligible pattern. That stronger forms of responsiveness to reasons are not required for responsibility under their theory is compatible with the point made in the main text here. There are, of course, strong forms of responsiveness to reasons that are not enjoyed by addicts. However, there is little reason to think they are needed for full responsibility for wrong doing.

(8) . See Gideon Yaffe (2011), "Lowering the Bar for Addicts" in *Addiction and Responsibility*, George Graham and Jeffrey Poland (Eds.), Cambridge, MA: MIT Press, and Gideon Yaffe, "Are Addicts Akratic? Interpreting the Neuroscience of Reward" in *Addiction and Self-Control*, Neil Levy (Ed.), Oxford: Oxford University Press (forthcoming).

(9) . For a discussion of the way in which addicts' tendencies to discount the future hyperbolically cause them to experience temporary preference shifts, see George Ainslie (2001), *Break down of Will*, Cambridge: Cambridge University Press.

(10) . Cf. Stephen Darwall (2006), *The Second-Person Standpoint: Morality, Respect, and Accountability*, Cambridge, MA: Harvard University Press; Susan Wolf (1987), "Insanity and the

Metaphysics of Responsibility” in *Responsibility, Character, and the Emotions*, Ferdinand Schoeman (Ed.), Cambridge: Cambridge University Press.